



T.C.

İSTANBUL ALTINBAŞ UNIVERSITY  
ELECTRICAL AND COMPUTER ENGINEERING

**PREDICTING BREAST CANCER USING  
GRADIENT BOOSTING MACHINE**

Saher Imad Abed

Master Of Science

Thesis Supervisor

Asst. Prof. Dr. Sefer Kurnaz

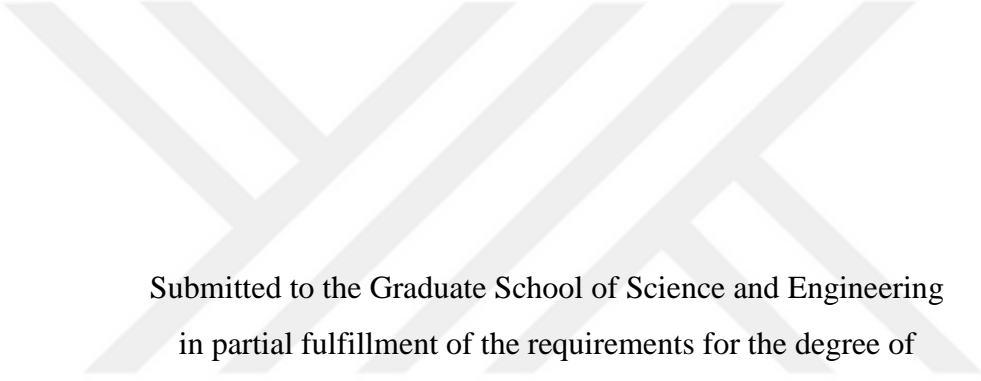
Istanbul, 2019

# **PREDICTING BREAST CANCER USING GRADIENT BOOSTING MACHINE**

By

Sahr Imad Abed

Electrical and Computer Engineering



Submitted to the Graduate School of Science and Engineering  
in partial fulfillment of the requirements for the degree of  
Master of Science

ALTINBAŞ UNIVERSITY

2019

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

---

Asst. Prof. Dr. Sefer KURNAZ  
Supervisor

Examining Committee Members (first name belongs to the chairperson of the jury and the second name belongs to supervisor)

|                              |   |       |
|------------------------------|---|-------|
| Prof. Dr. Osman UCAN         | Electrical and Electronic<br>Engineering , Altinbaş<br>university | _____ |
| Asst. Prof. Dr. Safer KURNAZ | Electrical and Electronic<br>Engineering , Altinbaş<br>university | _____ |
| Asst. Prof. Dr. Zeynep ALTAN | Engineering and Architecture<br>Faculty, Beykent University       | _____ |

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

---

Asst. Prof.Dr. Çağatay AYDIN  
Head of Department

Approval Date of Graduate School of  
Science and Engineering: \_\_\_\_/\_\_\_\_/\_\_\_\_

---

Assoc. Prof.Dr. Oğuz BAYAT  
Director

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Saher Imad Abed

## DEDICATION

I dedicate this thesis to my family, classmates and all my friends for the support and encouragement throughout my education and life. Special dedication goes to my supervisor Dr. Sefer Kurnaz, my sisters and my mother for their support and prayers during my research work.



## ACKNOWLEDGEMENTS

I might want to offer my thanks to every one of the individuals who have upheld me all through the regularly extended periods of this voyage. I might want to thank my advisor, Assist. Prof. Dr. Sefer Kurnaz for being my compass notwithstanding when I believed I was lost and being in extraordinary part in charge of the zenith of this work. I might likewise want to thank my supervisors for their supportive exhortation, which incredibly enhanced the nature of this work. Finally, I thank this institution for hosting me during these years, securely earning its place as my home.

I would also like to thank my family who has always unconditionally supported and motivated me throughout this whole process, and to whom I owe my every achievement. Also to my friends and loved ones who have always been by my side and kept me going through the hardest times.

## ABSTRACT

### Predicting Breast Cancer Using Gradient Boosting Machine

Abed, Sahr Imad

M.Sc. Electrical and Computer Engineering, Altınbaş University

Supervisor: Asst. Prof. Dr. Safer Kurnaz

Date: March /2019

Pages: 63

Breast Cancer is the most fatal 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 breast cancer, prognosis is related to the patterns of prediction. Cancer Prediction describes cancer that reappears after treatment, and in the specific case of breast cancer, prediction is very common, being experienced by about one third of patients after initial diagnosis. Therefore, establishing the patterns of prediction 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. Gradient Boosting is a powerful machine learning algorithm founded on the idea that combining the labels of many 'weak' classifiers or learners translates to a strong robust one to predict the breast cancer. Boosting is a greedy algorithm that fits adaptive models by sequentially adding these base learners to weighted data where difficult to classify points are weighted more heavily. Experts claim that gradient boosting is the best off-the-shelf classifier developed so far to detect and predict the Breast Cancer. As we can see from the above versions of boosting, a unique boosting algorithm can be derived for each loss function and its performance can vary depending on which base learner. We can derive a generic version of boosting called gradient boosting for the identification, detection, recognition and prediction of breast cancer.

**Keywords:** breast cancer, classification, machine learning, data mining, gradient boosting machine, prediction based system

# ÖZET

## GRADYAN ARTTIRMA MAKİNESİNİ KULLANARAK MEME KANSERİ TAHMİNİ

Abed, Sahr Imad

Yüksek Lisans, Elektrik ve Elektronik Mühendisliği, Altınbaş Üniversitesi

Danışman: Yrd. Prof. Dr. Safer Kurnaz

Tarih: Mart / 2019

Sayfalar: 63

Meme Kanseri, bu gibi yüksek ölüm oranlarına sahip en ölümcül hastalıklardır, sağkalm tahmini önemli bir rol üstlenir, çünkü klinisyenler her hastanın prognozunu ve denenecek ilgili tedavileri daha iyi tanımlamaya yardımcı olur. Özellikle meme kanseri için prognoz, tahmin şekilleriyle ilgilidir. Kanser Tahmini, tedaviden sonra ortaya çıkan kanseri ve spesifik meme kanseri durumunda, ilk teşhisten sonra hastaların yaklaşık üçte biri tarafından tecrübe edilen tahmin çok yaygındır. Bu nedenle, öngörü kalıplarını oluşturmak, bu patolojinin klinik davranışını doğru bir şekilde tahmin etmek için çok önemli bir görevdir. Bu, istenmeyen aşırı tedavi ve istenmeyen komplikasyonlardan kaçınarak hastalar için daha kişiselleştirilmiş bir tedavi sağlar. Gradient Boost, birçok "zayıf" sınıflandırıcı veya öğrencinin etiketlerinin birleştirilmesinin göğüs kanserini öngörmek için güçlü bir sağlam kelimeye çevrildiği fikrine dayanan güçlü bir makine öğrenme algoritmasıdır. Yükseltme, bu temel öğrenenleri sırayla sınıflandırmak zor noktaların daha ağır olduğu ağırlıklı verilere ekleyerek uyarlamalı modellere uyan açgözlü bir algoritmadır. Uzmanlar, gradyan artırmanın, Meme Kanseri'ni tespit etmek ve tahmin etmek için şimdiye kadar geliştirilen en iyi kullanıma hazır sınıflandırıcı olduğunu iddia ediyor. Yukarıdaki yükseltme sürümlerinden görebileceğimiz gibi, her kayıp işlevi için benzersiz bir yükseltme algoritması türetilir ve performansı, hangi temel öğrenciye bağlı olarak değişebilir. Meme kanserinin tanımlanması, tespiti, tanınması ve öngörülmesi için gradyan artırma adı verilen takviye işleminin genel bir versiyonunu türetebiliriz.

**Anahtar Kelimeler:** meme kanseri, sınıflandırma, makine öğrenmesi, veri madenciliği, gradyan artırma makinesi, tahmin tabanlı sistem



# TABLE OF CONTENT

Pages

|  |           |
|--|-----------|
| ABSTRACT.....  | vi        |
| TABLE OF CONTENT .....   | ix        |
| LIST OF TABLES .....   | xi        |
| LIST OF FIGURES.....   | xii       |
| LIST OF ABBREVIATION .....   | xiii      |
| <b>1. INTRODUCTION .....</b>   | <b>1</b>  |
| 1.1. CONTEXT .....   | 1         |
| 1.2. SOCPE OF THESIS.....  | 2         |
| 1.3. PLANNING .....  | 3         |
| 1.3.1. Familiarization with Breast Cancer and Gradient Boosting<br>Machine ..... | 3         |
| 1.3.2. Literature review of Boosting Machine .....                               | 3         |
| 1.3.3. Data gathering and analysis for Breast Cancer .....                       | 4         |
| 1.3.4. Defining approaches for Gradient Boosting .....                           | 4         |
| 1.3.5. Missing Data handling.....  | 4         |
| 1.3.6. Implementing pattern recognition techniques using GBM....                 | 4         |
| 1.3.7. Results: comparison and conclusions.....                                  | 4         |
| 1.3.8. Dissemination of Results.....   | 5         |
| 1.4. RISK ANALYSIS AND MITIATION .....   | 5         |
| 1.4.2. Techniques not applied in the medical context for boosting..              | 6         |
| 1.4.3. Dataset not available.....  | 6         |
| 1.4.4. Dataset requiring preprocessing.....                                      | 6         |
| 1.4.5. Dataset delivery is delayed .....   | 7         |
| 1.4.6. Algorithms are too complex.....   | 7         |
| 1.4.7. Algorithms' code is not available.....                                    | 8         |
| 1.5. DOCUMENT STRUCTRE .....   | 8         |
| 1.6. OUTLINE .....   | 9         |
| <b>2. BACKGROUND .....</b>   | <b>10</b> |
| 2.1. BREST CANCER.....   | 10        |
| 2.2. DATA MINING .....   | 12        |
| 2.2.1. <i>k</i> -Nearest Neighbors.....  | 14        |
| 2.2.2. Artificial Neural Networks .....  | 14        |
| 2.2.3. Gradient Boosting Machine.....  | 14        |
| 2.3. RECURRENCE SITES.....   | 15        |
| 2.4. DATA MINING APPROACHED .....  | 17        |
| 2.5. APPROACH OVERVIEW .....   | 21        |
| 2.6. DATA NORMALIZATION .....  | 22        |

|   |                                     |
|---|-------------------------------------|
| <b>3. METHODOLOGY .....</b>                         | <b>23</b>                           |
| 3.1. GRADIENT BOOSTING.....                         | 23                                  |
| 3.1.1. Discrete AdaBoost.....                       | 24                                  |
| 3.1.2. Understanding Gradient Boosting Errors ..... | 25                                  |
| 3.1.3. Weighted Training Errors.....                | 25                                  |
| 3.2. ENSEMBLE PREDICTION.....                       | 26                                  |
| 3.3. MULTICLASS CLASSIFACTION .....                 | 27                                  |
| 3.3.1. Gradient minimization.....                   | 27                                  |
| 2.4. GRADIENT BOOSTING ALGORTHIM.....               | 28                                  |
| 3.4. NOISE IDENTIFICATION FOR PREDECTION .....      | 30                                  |
| <b>4. SOLUTION .....</b>                            | <b>32</b>                           |
| 4.1. DATASET CHARACTERIZATUIN .....                 | 32                                  |
| 4.1.1. Inputs .....                                 | 32                                  |
| 4.1.2. Outputs.....                                 | 35                                  |
| 4.2. MISSING DATA HANDLING.....                     | 35                                  |
| 4.2.1 Missing Data simulation .....                 | 35                                  |
| 4.2.2 Validation.....                               | 37                                  |
| 4.3. CLASSIFICATION.....                            | 38                                  |
| 4.3.1 Classification Algorithms.....                | 38                                  |
| 4.3.2 Validation.....                               | 39                                  |
| 4.4. CONCLUSION .....                               | 40                                  |
| <b>5. RESULTS .....</b>                             | <b>41</b>                           |
| 5.1 FEATURESELECTION.....                           | 41                                  |
| 5.2 IMPUTATION.....                                 | 41                                  |
| 5.2.1 Imputations by algorithm .....                | 41                                  |
| 5.2.2 Final datasets for classification.....        | 44                                  |
| 5.3 CLASSIFICATION.....                             | 44                                  |
| 5.3.1. Distribution of Classes .....                | 49                                  |
| <b>6. DISCUSSION .....</b>                          | <b>53</b>                           |
| 6.1. DISCSSION.....                                 | <b>Error! Bookmark not defined.</b> |
| 6.2. DATA ANALYSIS .....                            | 54                                  |
| <b>7. CONCLUSION.....</b>                           | <b>56</b>                           |
| 7.1. CONCLUSION.....                                | <b>Error! Bookmark not defined.</b> |
| 7.2. FUTURE WORK.....                               | 57                                  |
| <b>REFERENCES .....</b>                             | <b>58</b>                           |

## LIST OF TABLES

|  | <u>pages</u> |
|--|--------------|
| <b>Table 4.1:</b> Provides the diagnosis between malignant and benign based on the attributes of radius, texture and perimeter mean to classify the output metrics ..... | 34           |
| <b>Table 4.2:</b> Outputs respective to prediction sites .....   | 35           |
| <b>Table 5.1:</b> Imputation results for Benign and Malignant .....  | 42           |
| <b>Table 5.2:</b> Second iteration of ANN Imputation .....   | 42           |
| <b>Table 5.3:</b> First iteration of GBM Imputation .....  | 43           |
| <b>Table 5.4:</b> Second iteration of GBM Imputation.....  | 43           |
| <b>Table 5.5:</b> Third iteration of GBM Imputation .....  | 43           |
| <b>Table 5.6:</b> Complete datasets resulting from the use of GBM prediction and classification for tumors .....   | 44           |
| <b>Table 5.7:</b> Description of methodology and results. ....   | 52           |

## LIST OF FIGURES

|   | <u>Pages</u> |
|---|--------------|
| <b>Figure 2.1:</b> Depiction of the Supervised Learning process .....   | 13           |
| <b>Figure 3.1:</b> Weighted training error $\epsilon t$ for the boosting algorithm on an artificial set of iterations to detect and predict the Breast Cancer. .... | 26           |
| <b>Figure 3.2:</b> Ensemble Prediction with exponential upper bound for the boosting algorithm on an artificial set. ....   | 27           |
| <b>Figure 3.3:</b> Training and testing the dataset using GBM algorithm on a pre-trained model. ....  | 29           |
| <b>Figure 3.4:</b> A schematic of discarded vs mislabeled points for a noise identification .....   | 31           |
| <b>Figure 4.1:</b> Gradient Boosting Machine Model for predicting the labelled data in classes.   | 38           |
| <b>Figure 4.2:</b> Example of a ROC curve .....   | 40           |
| <b>Figure 5.1:</b> Radius, Perimeter and Area have strong positive correlation .....  | 45           |
| <b>Figure 5.2:</b> Radius have a positive correlation with Concave Points .....   | 45           |
| <b>Figure 5.3:</b> Compactness, Concavity and Concave Points have strong positive correlation   | 46           |
| <b>Figure 5.4:</b> Fractal Dimension have some negative correlation with Radius, Perimeter and Area .....   | 47           |
| <b>Figure 5.5:</b> Fractal Dimension have some negative correlation with Radius, Perimeter and Area .....   | 48           |
| <b>Figure 5.6:</b> Total number of tumors being classified using the gradient boosting machine with count given vertically .....                                    | 49           |
| <b>Figure 5.7:</b> Selected features being mapped on the surface grounds of 32-classes in the dataset. ....   | 50           |
| <b>Figure 5.8:</b> The prediction of malignant in green and benignant in red based on trained model using gradient boosting machine. ....                           | 51           |

## LIST OF ABBREVIATION

|        |   |
|--------|---|
| AUC    | Area Under the receiver operating characteristic Curve        |
| AvAp   | Average Accuracy in percentage                                |
| AvAU   | Average AUC   |
| BC     | Breast Cancer   |
| CA15-3 | Cancer Antigen 15-3   |
| DM     | Data Mining   |
| ER     | Estrogen Receptor (protein)                                   |
| HER2   | Human Epidermal growth factor Receptor 2 (protein)            |
| GBM    | Gradient Boosting Machine                                     |
| IHC    | Immuno-Histo-Chemistry (process for protein detection)        |
| L1QP   | L1 soft-margin minimization by quadratic programming          |
| MD     | Missing Data  |
| ML     | Machine Learning  |
| PR     | Progesterone Receptor (protein)                               |
| RBF    | Radial Basis Function   |
| ROC    | Receiver Operating Characteristic                             |
| SEER   | Surveillance, Epidemiology, and End Results                   |
| SMO    | Sequential Minimal Optimization (GBM routine)                 |
| WEKA   | Waikato Environment for Knowledge Analysis (programming tool) |
| WHO    | World Health Organization                                     |

# 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 Breast Cancer is rare: it represents less than 1% of all cases [2]. Further references to Breast Cancer will pertain to female Breast Cancer except where noted, since it is what this work will focus in.

World follows these global trends, with Breast Cancer 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 World [1]. According to WHO (World Health Organization) projections, these number are expected to rise, with 1620 deaths by Breast Cancer 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 prediction [4]. Cancer Prediction (or Re-lapse) describes cancer that reappears after treatment, and in the specific case of BC, prediction is very common, being experienced by about one third of patients after initial diagnosis [4]. Therefore, establishing the patterns of prediction 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 Breast Cancer in the last couple of decades, the underlying processes of prediction have not yet been completely understood [5]. Encompassed in this reality, this work conducts a data-driven research, attempting to construct a model of prediction for patients with this condition. As detailed in the following section, our goal is to study the prognostic factors that define female Breast Cancer prediction, clarify the correlation between such factors and relapse patterns, and lastly, to provide a model to predict prediction for a particular patient, based on her personal characteristics as well as her tumor expression.

## **1.2. SOCPE OF THESIS**

Our thesis aims to construct a model of metastatic gradient boosting machine for Breast Cancer prediction. This is achieved by examining the behavior of Breast Cancer relapses, in terms of the localization of the tumor and its other features. The primary goals of this work are the following:

### **A. Evaluate the pattern of metastatic dissemination in patients with Breast Cancer tumors.**

The first objective is to understand how the relapses are physically distributed, and their respective characteristics. Breast Cancer prognosis is related to prediction, 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 prediction metastases.

### **B. Establish the relation between the patterns of metastatic proliferation, patient's characteristics and Breast Cancer subtypes using the dataset.**

After analyzing the metastatic spread of Breast Cancer, it will be measured the correlation between these data and the characteristics of the patients and their tumors. Breast Cancer 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 Breast Cancer prediction.

### **C. Build a model of Breast Cancer prediction using GBM.**

This work intends to define a prediction 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 on this thesis. The time expected to complete each task can be compared to the real period spent to complete it for predicting the Breast Cancer.

##### **1.3.1. Familiarization with Breast Cancer and Gradient Boosting Machine**

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 Breast Cancer, and the related terminology, and its prediction, 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.

##### **1.3.2. Literature review of Boosting Machine**

There has been an increase in awareness regarding Breast Cancer. 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 Breast Cancer relapses is 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 Breast Cancer prediction, including several studies analyzing the metastatic behavior of Breast Cancer.

### **1.3.3. Data gathering and analysis for Breast Cancer**

The data used in this work were received, 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.

### **1.3.4. Defining approaches for Gradient Boosting**

Due to the complex nature of Breast Cancer, and cancer tumors in general, we need to determine the aspects in which we will focus our analyses. As such, Breast Cancer prediction site was chosen as the primary splitting point, which guided the division of sections in this thesis.

### **1.3.5. 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 Breast Cancer 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.

### **1.3.6. Implementing pattern recognition techniques using GBM**

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.

### **1.3.7. 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.

### **1.3.8. Dissemination of Results**

The thesis' Final Report is written in this task, including all of the work developed during the year on Gradient Boosting Machine. Moreover, a scientific article is also be produced.

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

## **1.4. RISK ANALYSIS AND MITIATION**

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 a structured manner.

The first two tasks are “Familiarization with Breast Cancer” 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:

### **1.4.1. Breast Cancer prediction papers are too specific using GBM**

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 Breast Cancer, the same is not verified when dealing with its prediction 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 Breast Cancer prediction, it is important to read about the primary disease itself.

#### **1.4.2. Techniques not applied in the medical context for boosting.**

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

#### **1.4.3. 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 Breast Cancer, 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].

#### **1.4.4. 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.

#### **1.4.5. 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 also includes the implementation phase. Regarding this step, the following risks were found:

#### **1.4.6. 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.

#### **1.4.7. 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.

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:

### **1.5. DOCUMENT STRUCTRE**

The following chapters show the remainder of our work: Chapter 2 contains more detailed information regarding the main topics of this thesis: Breast Cancer as a pathology, GBM techniques, and evaluation metrics of classification systems. Chapter 3 reveals an analysis of recent related literature regarding Breast Cancer Prediction. This State of Art is divided into two sections, focusing on the prediction sites on one hand, and GBM 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.

## 1.6. OUTLINE

The goal of this thesis is to further the understanding of boosting algorithm of machine learning in predicting the Breast Cancer via a volume based approach by developing an understanding for how the weak learners are chosen to affect noisy points as opposed to non-noisy points. The findings further tie this approach to our understanding of boosting in a margin sense and how these margins change with each successive iteration on noisy training examples. An important component of the analysis centers around the ability to filter noisy data, so the efficacy of the noise filter techniques in the literature is compared with a noise identification scheme based on examining the decision volume around each training example and apply this methodology to real-world datasets for validation.

This thesis is organized as follows:

- Chapter 2 contains all related work regarding gradient boosting machine with other machine learning algorithms for comparison.
- Chapter 3 describes the methodology and dataset which we will employ.
- Chapter 4 summarizes the results of the experiments.
- Chapter 5 summarizes the results of the experiments.
- Chapter 6 discusses the experiments and results in terms of gradient boosting.
- Chapter 7 concludes the thesis and delves into the future directions of this study.

## 2. BACKGROUND

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 breast cancer 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. Breast Cancer 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 Breast Cancer 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 Breast Cancer, detecting cancer much before any symptoms show up. External signs of Breast Cancer 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 Breast Cancer 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 lobular carcinomas start in the lobules, representing about 10% of invasive Breast Cancer.

Breast Cancer 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 process 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 Breast Cancer tumors.

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

- Luminal: ER<sup>+</sup> or PR<sup>+</sup> (at least one of them) and HER<sup>-</sup>
- HER2-enriched: HER<sup>+</sup>
- Triple-negative: ER<sup>-</sup>, PR<sup>-</sup> and HER<sup>-</sup>

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

Breast Cancer 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 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

Adjuvant therapy for Breast Cancer 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



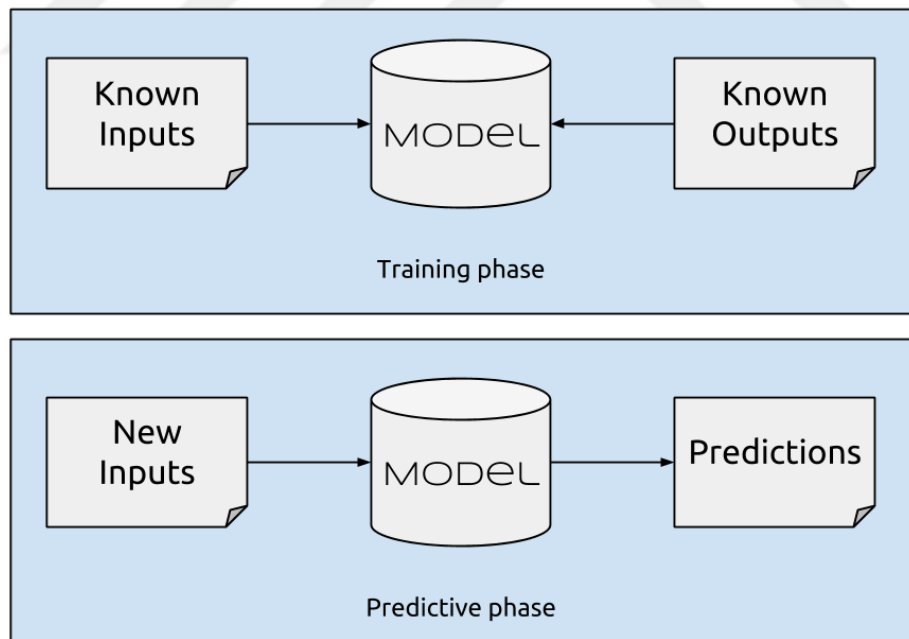
Breast Cancer cells from receiving the hormone estrogen; Tamoxifen, for example, blocks estrogen's activity in the body. Trastuzumab is a targeted drug, focusing on cells that ever express HER2; Radiation therapy is usually given after breast-conserving surgery and may be given after a mastectomy (it is a local therapy, while the others are systemic therapies, because they travel to the whole body through the bloodstream). Neo adjuvant 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 Breast Cancer patients, including those with prediction.

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 Breast Cancer 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.

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

### **2.2.3. Gradient Boosting Machine**

Data are represented by a set of feature vectors. When the data have two classes,

GBM can try to divide them with a hyper-plane. To do this, the algorithm projects the examples in a higher-dimensional space, simplifying the problem of creating a division. GBM 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 GBM 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. Theoretically, GBM’s 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, 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

### **2.3. RECURRENCE SITES**

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

There has been significant progress in the characterization of Breast Cancer.

However, it is often still difficult to accurately predict its behavior [17]. In luminal like disease (hormonal receptor positive, HR<sup>+</sup>), this is especially. In comparison, the value associated with prediction is of around 20 000, obtained when including the results for any combination of the terms “prediction(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 prediction 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 Breast Cancer 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 prediction in early stage breast cancer differs according to the metastatic site [23], for example. The tests used were  $\chi^2$  (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 prediction, and those who hadn’t died at cutoff date). Having established the cumulative curves across Breast Cancer subtypes, Gray’s test was the choice to compare them, testing them for statistically significant differences. Survival (from initial and prediction 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 Breast Cancer 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 prediction 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.

#### **2.4. DATA MINING APPROACHED**

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 prediction.

In 2017, 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 remaining 90% (802) had no evidence of it (10% rate of prediction). Since the two classes were imbalanced, 6 different sub-datasets were created, each with 148 relapse-free patients

and all of the 85 with prediction (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  $\approx 68.3\%$ ). To properly evaluate the techniques, 50 runs were conducted for each sub-dataset, splitting into different partitions of training set ( $n=155$ ) and test set ( $n=78$ ). Averaging the accuracy of the 300 runs ( $50 \times 6$ ), the second best value was achieved by FOCL ( $\approx 66.4\%$ ). However, this was the only metric used, which does not allow a full comparison of performance.

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 prediction 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. Using a 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 prediction. 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 prediction. 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 Breast Cancer relapse using Gradient Boosting Machine. 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 prediction, 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%).

In 2007, the same authors applied once again Gradient Boosting Machine to predict prediction in Breast Cancer [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 Gradient Boosting Machine 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, Gradient Boosting Machine had lower AUC values, but the differences were not statistically significant. In terms of predictive power of Prediction, 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 Gradient Boosting Machine, 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 prediction 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 Gradient Boosting Machine and several Rule Inducers, among others, although these algorithms were only used with one configuration. One of the goals was to predict whether a Breast Cancer patient would develop prediction 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. 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. As feature selection, medical consulting resulted in 13 attributes being selected as inputs. The algorithms used included ANN and GBM, with four variants of the latter. Dividing the dataset into training and test partitions (80%/20%) showed that C5 Gradient Boosting Machine GBM had the best accuracy (71%),

but ANN provided better predictive power for the Prediction class (78%, higher than 72% by C5 GBM), 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 prediction in a public Breast Cancer 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. 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.

## **2.5. APPROACH OVERVIEW**

As can be seen in the two previous sections, there have been many studies regarding the pathways of Breast Cancer prediction. The main purpose of the first section was to systematize the previous work in this area using AdaBoost or Gradient Boosting Machine which is a well-known supervised machine learning algorithm. The authors of these research studies use statistical algorithms of GBM 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 prediction site is presented.

Machine Learning algorithms have also been applied to the study of Breast Cancer prediction, with the capacity of unveiling information hidden in the data, generalizing from its underlying patterns. We use GBM for 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 of breast cancer prediction, and there have been interesting developments in this field of gradient boosting machine for prediction.

## **2.6. DATA NORMALIZATION**

Normalization provides the fundamental knowledge required to understand the different steps of this thesis. It covers both the algorithmic and technological points of view. The terminology and core concepts of Breast Cancer 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 Breast Cancer prediction using gradient boosting machine and other algorithms. It is more difficult for someone outside the medical community to understand, but the collaboration has the potential to be very rewarding. The data mining section explains the theory of the computational aspects of this thesis.

### 3. METHODOLOGY

This chapter will present a selection of research papers about Breast Cancer Prediction, divided into two categories. While Section 3.1 contains articles with a focus on the gradient boosting, Section 3.2 shows the use of predictive machine learning in Breast Cancer prediction using Ada-boost. 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 prediction as an atomic event. Nevertheless, these show some of the applications of Data Mining techniques to the study of prediction of malignant breast tumors. Finally, a brief conclusion of this review work will be provided.

#### 3.1. GRADIENT BOOSTING

Gradient Boosting is a powerful machine learning algorithm founded on the idea that combining the labels of many ‘weak’ classifiers or learners translates to a strong robust one to predict the breast cancer. Boosting is a greedy algorithm that fits adaptive models by sequentially adding these base learners to weighted data where difficult to classify points are weighted more heavily. Experts claim that boosting is the best off-the-shelf classifier developed so far to detect and predict the Breast Cancer.

The goal of boosting is to minimize the function to predict the cancerous cells.

$$\min = \sum_{i=1}^n \text{Loss}(y_i, h(x_i)) \quad (3.1)$$

In a stage-wise manner where many different loss functions can be used to detect and predict the breast cancer. At each iteration, the goal of minimizing the above problem is approached in a stage-wise manner where a new classifier is added each time to predict the breast cancer. Since the previous parameters cannot be changed, we call this approach forward stage-wise additive modeling to detect and predict the Breast Cancer. The primary tuning parameter in forward stage-wise additive modeling is the number of iterations. This

parameter can be tuned via a validation set where the parameter can be chosen to be the point where the performance begins to decrease called early stopping to detect and predict the Breast Cancer. Alternative parameters such as AIC or BIC can also be used. Another technique for achieving better generalization performance is to enforce a learning rate on each update making the first few iterations more ‘important’ than the last few. This technique is typically referred to as shrinkage to predict the breast cancer. In binary classification problems, it is natural to use 0-1 loss; however, since 0-1 loss is not differentiable different boosting techniques may use log loss or exponential loss as a convex upper bound for 0-1 loss to detect and predict the Breast Cancer.

### 3.1.1. Discrete AdaBoost

The most popular boosting algorithm, Adaboost, was developed by Freund and Schapire (2017) [10] solving many of the practical drawbacks of earlier boosting methods utilizing an exponential loss function. The Adaboost algorithm takes a training dataset  $S_n = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$  as input with binary labels  $Y = \{-1, +1\}$  (although it can be extended to the multiclass case which we will cover in a later section). Ada-boost iteratively calls a series of base learners in a series of rounds  $t = 1 \dots T$ . Each new base learner  $h(x; \theta_t)$  improves upon the overall classification of the ‘committee’ of base learners by weighting the  $i^{th}$  training examples for round  $t$  denoted by  $W_t(i)$  to detect and predict the Breast Cancer. The weights are determined based on the classification performance of the ensemble of classifiers in previous iterations to predict the breast cancer. If the classifier misclassifies a training example, then the weight of the training example increases. Thus, the subsequent base learners focus on the examples that are hard to classify. After a base learner is chosen, Ada-boost chooses an importance weight  $\alpha_t$  for the classifier based on the error of the chosen classifier on the weighted training set to predict the breast cancer. Thus, as the error of iteration  $t$ ,  $\epsilon_t$  increases, then the importance weight  $\alpha_t$  decreases to detect and predict the Breast Cancer. Note that  $\alpha_t$  cannot be updated in subsequent rounds but can only be changed by choosing the same learner in a later boosting round. Thus, we do not require that they sum to 1 and can simply normalize them later on. The final hypothesis resulting from the Ada-boost algorithm is a weighted majority vote by the committee of base learners to detect and predict the Breast Cancer.

The final ensemble  $H_T(x)$  can be written as:

$$H_T(x) = \sum_{i=1}^T h(x; \theta_i) \quad (3.2)$$

Where  $h(x; \theta_t)$  is the base learner in boosting round  $t$  with parameters  $\theta_t$ . Ada-boost in particular trains its ensemble classifier based on an exponential loss function.

$$\text{Loss}(y, h(x)) = \exp(-yh(x)) \quad (3.3)$$

Where  $h$  is the classifier. The primary advantage of the exponential loss function is computational in its simplicity to predict the breast cancer. We choose the values of  $\alpha t$ ,  $\theta t$  by minimizing the loss function to predict the breast cancer.

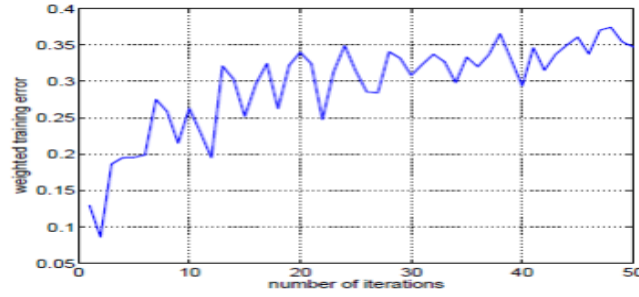
Note that since this derivative is negative, we can expect the training loss to decrease by adding our new base learner to predict the breast cancer. Now that we have all of the necessary components for the Ada-boost algorithm, we can now explicitly define the steps. We will only examine Discrete Ada-boost which uses decision stumps  $h(x; \theta) = \text{sign}(\theta tx + \theta t)$  as base learner (though there are other versions that use real-valued classifiers such as in Real Ada-boost). Often it has been found that classification trees make good base learners. Note that decision stumps are special cases of a classification tree with depth 1.

### 3.1.2. Understanding Gradient Boosting Errors

We describe the weighted training error  $\epsilon t$  as the weighted error of the  $t^{\text{th}}$  base learner with respect to the weights  $W_{t-1}(i)$  on the training set to detect and predict the Breast Cancer. We can also measure the performance of the classifier on the next iteration by taking its error with respect to  $(i)$ .

### 3.1.3. Weighted Training Errors

The weighted training error  $\epsilon t$  with respect to  $W_{t-1}(i)$  increases as more boosting iterations occur although it does not do so in a monotonic fashion as shown in Figure 2-1 to detect and predict the Breast Cancer. This matches our intuition since the weights increase for difficult to classify points making the weighted training error more difficult to minimize for each subsequent base learner.



**Figure 0.1:** Weighted training error  $\epsilon t$  for the boosting algorithm on an artificial set of iterations to detect and predict the Breast Cancer.

Implying that the weighted agreement of the predicted and true label on the updated weights is 0 (this is equivalent to random guessing). This property has strong implications as it implies that the base learner from the  $t^{th}$  boosting iteration will be useless for the next iteration. This prevents the same base learner from being chosen two iterations in a row. This makes sense as if you had the same base learner twice in a row, this is equivalent to choosing it once while summing their  $\alpha$  weights to detect and predict the Breast Cancer. This property ensures that weights are assigned efficiently. However, the same learner can appear in the future with respect to different weights since we never have a chance to go back and update previous  $\alpha t$ 's, so in order to tune them, we have to include the learner in a future iteration to detect and predict the Breast Cancer.

### 3.2. ENSEMBLE PREDICTION

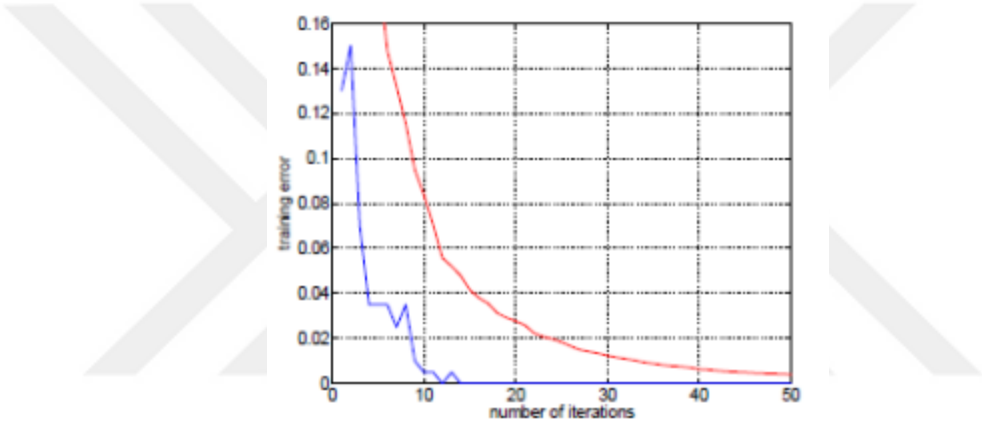
Now that we have examined how the error varies from iteration to iteration, it is important to understand how the error of the entire ensemble behaves for making a prediction on the cell of tumors in the breast of women. The ensemble training error does not decrease monotonically with each boosting iteration to predict the breast cancer. However, the exponential loss function which Gradient Boosting Algorithms chooses weights and base learners to sequentially optimize for does decrease monotonically with each boosting iteration. We can compute exactly how much the exponential loss decreases on each successive iteration by the equation given by to predict the cancerous cells in breast.

$$(\alpha t, \theta t) = \sum_{i=0}^n W t^{-1}(i) \exp(y_i \alpha(x_i; \theta t)) + (1 - \epsilon t) \exp(-\alpha t) + \epsilon t \exp(\alpha t) \quad (3.4)$$

Notice that the above expression  $(\alpha t, \theta t)$  is equivalent to the amount that we renormalize the

weights by and also the amount that the exponential loss decreases on each iteration. When  $\epsilon t < 1/2$ , then this value is  $< 1$ , so the exponential loss is ensured to decrease by a factor to predict the cancerous cells in breast of women. Thus, the overall ensemble prediction error for exponential loss ( $Ht$ ) after  $t$  iterations is simply a product of these normalization constants to predict the breast cancer.

Since the zero-one loss in computing ensemble training error is upper bounded by our exponential loss definition, then the ensemble training error is guaranteed to decrease the more iterations that occur. A plot of the ensemble training error (blue) and the exponential loss upper bound (red) can be seen in Figure 3.2



**Figure 0.2:** Ensemble Prediction with exponential upper bound for the boosting algorithm on an artificial set.

### 3.3. MULTICLASS CLASSIFICATION

Boosting has commonly been used in the two-class case. Common approaches for extending boosting to the multi-class classification problem is to reduce it to a series of two-class classification problems. In the case of Adaboost, a natural extension to the multiclass problem developed by Schapire and Freund [37] based on pseudo-loss. However, this extension of the binary problem has some drawbacks to predict the breast cancer. The importance weight  $\alpha$  requires each error  $\epsilon$  to be less than half with respect to the distribution it was trained on in order for the classifier to be properly boosted. However, when extending to the multiclass case, the random guessing rate is  $1/K$ , but the classifier still must perform better than  $1/2$  to avoid negative weights, which is much harder for a base learner to satisfy to predict the breast cancer by terming the malignant and benign cell.

#### 3.3.1. Gradient minimization



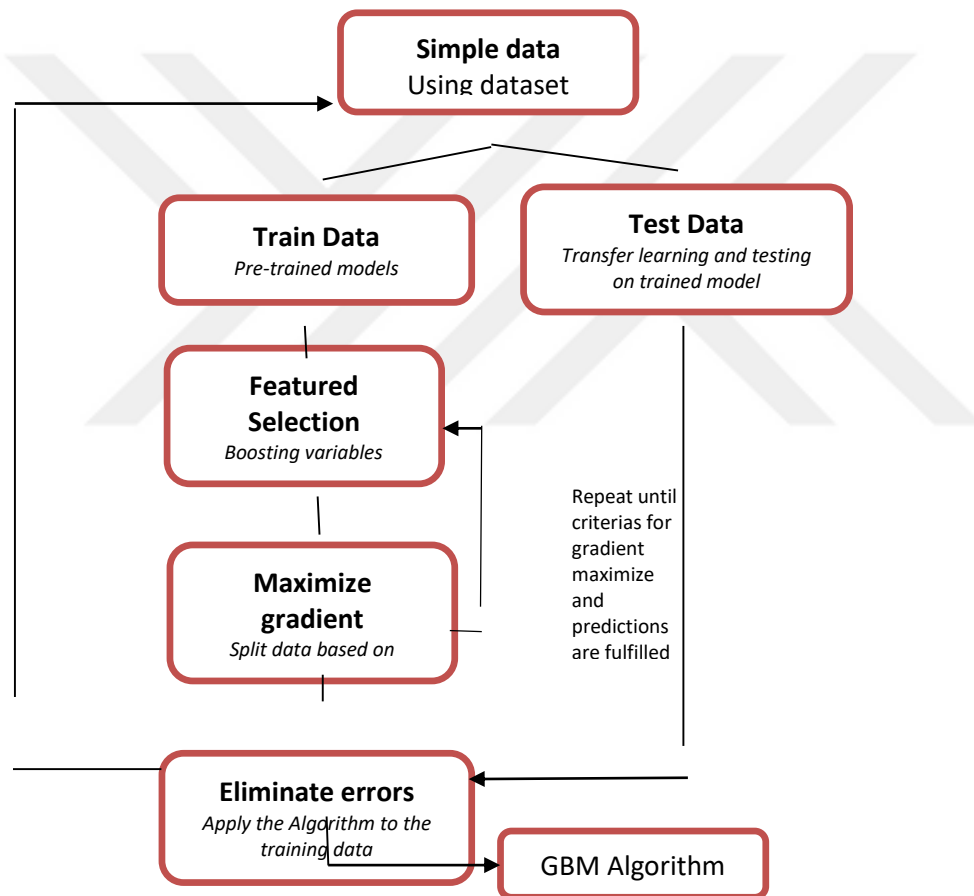
The problem with gradient boosting and utilizing an exponential loss function is that it puts a significant amount of weight on the misclassified or 'difficult' examples to predict the cancer in the breast. This causes the algorithm to be very sensitive to outliers (misclassified training examples). In addition to this sensitivity, since the exponential loss function is not the logarithm of any density functions, it is difficult to extract corresponding probability estimates from the value returned by  $(x)$  to predict the breast cancer by terming the malignant and benign cells into classes. A less harsh alternative is log loss which linearly concentrates on mistakes rather than exponentially. Such a boosting algorithm incorporating log loss seeks to minimize the tumor cells into malignant tumors. There are two types of single algorithm filters. One uses the same algorithm for both filtering and classifying. An example of this idea is fitting a dataset in regression analysis, removing the data points, and re-fitting on the modified dataset. Another common example of this approach is employed by John [18] in which a decision tree is created and subsequently pruned. The other approach in single algorithm filters is to use one algorithm for filtering and another one for classifying. The reason one might choose this approach over using the same algorithm for both is that some algorithms act as better filters while others may be better at classification.

#### **2.4. GRADIENT BOOSTING ALGORITHM**

As we can see from the above versions of boosting, a unique boosting algorithm can be derived for each loss function and its performance can vary depending on which base learner. We can derive a generic version of boosting called gradient boosting [42]. In this approach, we seek to find the function that minimizes the loss function written as:

$$h = \arg \min (h) \tag{3.5}$$

Where  $h$  is our function. We can view this as a gradient descent in the function space. We perform a stage-wise gradient descent prediction in a stage-wise fashion to get the gradient for the prediction of cancer in the current state of breast.



**Figure 0.3:** Training and testing the dataset using GBM algorithm on a pre-trained model.

The step-size of the functional gradient descent for the prediction of breast cancer is given above in equation (5). So far, this approach is not particularly useful because the function itself will most likely over-fit the training data. We would like to find a base learner that can approximate the negative gradient with the following function

$$\text{Loss}(y, (x)) = (y_i - H_{t-1}(x_i) - h(x_i; \theta_t)) \quad (3.6)$$

Furthermore, not only did it not over-fit the data, even after the training error reached zero, the generalization error continued to decrease with each subsequent iterations. There was significant interest behind understanding why such a complex hypothesis yielded extremely low error rates of detection of cancer cells in the breast.

Based on the work done by Bartlett [12] and Schapire et al. [17], in order to understand why boosting was resistant to over-fitting, the authors did not simply examine how boosting effected the training error (the number of misclassified examples), but rather examined the confidence of the classification. In their work, they modeled the confidence of classification

---

### Algorithm of Gradient Boosting Machine

---

1. Initialize  $h(x; \theta)$  as  $\theta = \arg \min_{\theta} \sum_{i=1}^n L(y_i, h(x_i; \theta))$
  2. For boosting stage  $t$ , compute the gradient  $(g_t)$
  3. Find the base learner  $h(x; \theta_t)$  that minimizes  $\sum_{i=1}^n (g_t(x_i) - h(x_i; \theta_t))$
  4. Now add the learner to the new prediction  $(x) = H_{t-1}(x) + \alpha h(x; \theta_t)$
- 

### 3.4. NOISE IDENTIFICATION FOR PREDECTION

In order to gain an understanding the behavior of gradient boosting methods on noisy data points, a method must be developed for identifying these noisy points on real-world data. Unfortunately, methodology for identifying noise still remain very rudimentary; otherwise, every machine learning problem would begin with filtering out all of the noisy data points from every dataset to predict the breast cancer. Separating the signal from the noise in itself is a hard problem and has strong effects on machine learning algorithms' ability to generalize [42]. The source of labeling errors can come in a variety of forms from subjective evaluation, data entry error, or simply inadequate information to predict the breast cancer.

For example, subjectivity can arise when a medical practitioner is attempting to classify disease severity. Another example in which there may be inadequate information is if a human records an image pixel based on color rather than the numeric input used by the algorithm to classify the data to predict the breast cancer. Wilson [39] first used the idea of noise filtering in which noisy points were identified to eliminate and improve classification

performance. Wilson employed a  $k$  nearest neighbor classifier to filter the dataset and fed the correctly classified points into a 1-nearest neighbor classifier to predict the breast cancer. Tomek [34] extended Wilson's algorithm for varying levels of  $k$ . Much further work was done on instance-based selection for the purpose of exemplar-based learning algorithms to predict the breast cancer.

Gamberger et al. [14] developed a noise identification algorithm that focused first on inconsistent data points, that is points with both labeling present for the same feature values to predict the breast cancer. After removing the inconsistent data, they transformed the features into binary values and removed features that most significantly reduced the number of literals needed to classify the data to predict the breast cancer. While many of the previous works have found noise detection methods for certain contexts in order to improve generalization and prediction, this does not fall in accordance with this study to predict the breast cancer.



**Figure 0.4:** A schematic of discarded vs mislabeled points for a noise identification

Filter in the breast for making prediction based on labels of data.

The danger of automatically flagging points that are difficult to correctly classify as noisy data is that they could be an exception to the rule rather than a noisy data point to predict the breast cancer. An important question is to determine a method for differentiating between exceptions and noise to predict the breast cancer. Guyon et al. [16] developed an information criterion to determine how typical a data point was, but since it was an on-line algorithm, it was sensitive to ordering. Srinivasan et al. [30] utilized an information theory-based method to separate exceptions and noise in the context of logical theory to predict the breast cancer. Oka et al. [25] developed methodology for learning generalizations and exceptions to the rule separately by separately noting which data points were correctly and

incorrectly classified in the neighborhood around each training example. Their algorithm for differentiating the noise from the exceptions rested upon a user input which made sure the classification rate passed the threshold to predict and detect the breast cancer and tumors.

## 4. SOLUTION

This chapter presents the implementation and solution used in this thesis. The first section describes the dataset characterization used in this work, while the other two present implementation details, inputs, outputs, validations, 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 Wisconsin. The study population is composed of female patients, older than 18 years of age, with breast carcinoma histologically confirmed in all of these patients. To protect the confidentiality of the patients, we never had access to their names, using an ID (IPI number) as distinguishable identifier. From a database with 274 patients, two of them did not contain the necessary information about prediction. Those were removed immediately, leaving a final cohort with a total of 272 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:

- 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
- Patients in this study have disease of either stage *I*, *II* or *III*;

| <b>Diagnosis</b> | <b>Radius Mean</b> | <b>Texture Mean</b> | <b>Perimeter Mean</b> |
|------------------|--------------------|---------------------|-----------------------|
| M                | 17.99              | 10.38               | 122.8                 |
| M                | 20.57              | 17.77               | 132.9                 |
| M                | 19.69              | 21.25               | 130                   |
| M                | 11.42              | 20.38               | 77.58                 |
| M                | 20.29              | 14.34               | 135.1                 |
| B                | 12.45              | 15.7                | 82.57                 |
| B                | 18.25              | 19.98               | 119.6                 |
| M                | 13.71              | 20.83               | 90.2                  |
| M                | 13                 | 21.82               | 87.5                  |
| M                | 12.46              | 24.04               | 83.97                 |
| M                | 16.02              | 23.24               | 102.7                 |
| B                | 15.78              | 17.89               | 103.6                 |
| B                | 19.17              | 24.8                | 132.4                 |
| M                | 15.85              | 23.95               | 103.7                 |
| M                | 13.73              | 22.61               | 93.6                  |
| M                | 14.54              | 27.54               | 96.73                 |
| M                | 14.68              | 20.13               | 94.74                 |
| B                | 16.13              | 20.68               | 108.1                 |
| M                | 19.81              | 22.15               | 130                   |
| B                | 13.54              | 14.36               | 87.46                 |
| B                | 13.08              | 15.71               | 85.63                 |

|   |       |       |       |
|---|-------|-------|-------|
| B | 9.504 | 12.44 | 60.34 |
| M | 15.34 | 14.26 | 102.5 |
| M | 21.16 | 23.04 | 137.2 |
| B | 16.65 | 21.38 | 110   |
| M | 17.14 | 16.4  | 116   |

- ER, PR and HER2 expression were determined via IHC;

**Table 4.1:** Provides the diagnosis between malignant and benign based on the attributes of radius, texture and perimeter mean to classify the output metrics



### 4.1.2. Outputs

Table 4.2 shows the output variables. Each of the variables refer to a single location, with exception of “Benign” and “Malignant”, 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 smoothness in the breast tissues.

**Table 4.2:** Outputs respective to prediction sites

| Variable  | Variable type | Smoothness |
|-----------|---------------|------------|
| Malignant | binary        | 0.1184     |
| Benign    | binary        | 0.0 474    |
| Malignant | binary        | 0.1096     |
| Benign    | binary        | 0.1425     |
| Malignant | binary        | 0.1003     |
| Benign    | binary        | 0.1278     |

0.09463

0.1189

## 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.

$$(m + 1)^v - 1 = 9^{12} - 1 = 282429536480$$

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 and predict the breast cancer. 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, GBM 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 GBM algorithm, the first iteration used only five values for  $k$  (number of neighbors), but all possible values for the distance. Afterwards, the same was done for ANN, GBM 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.

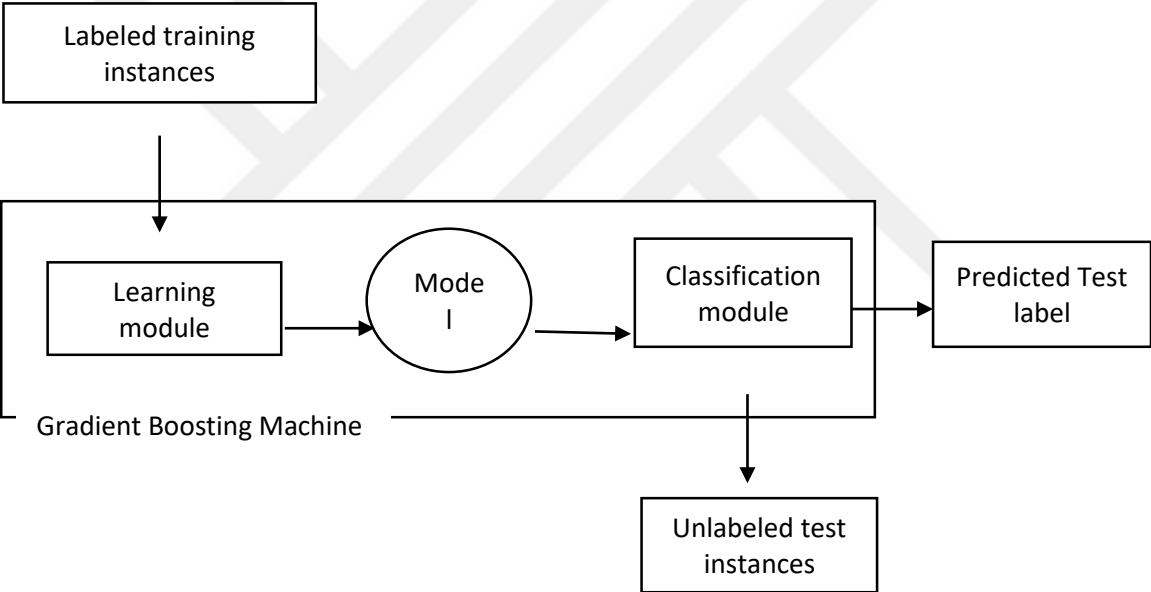
The choice for metric, due to its simplicity, is Accuracy, given by the following formula:

$$Accuracy = (Total\ Correct) / (Total\ Values\ to\ Impute)$$

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 used in gradient boosting machine (GBM) 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.



**Figure 4.1:** Gradient Boosting Machine Model for predicting the labelled data in classes.

#### 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 used in gradient boosting machine (GBM). For more information, see Figure 4.3. Besides the classifiers used in imputation, GBM 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 [62,71].

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 if three quarters of the outputs belong to the negative class, the model can have 75% accuracy using the algorithm just by assigning every patient to that class.

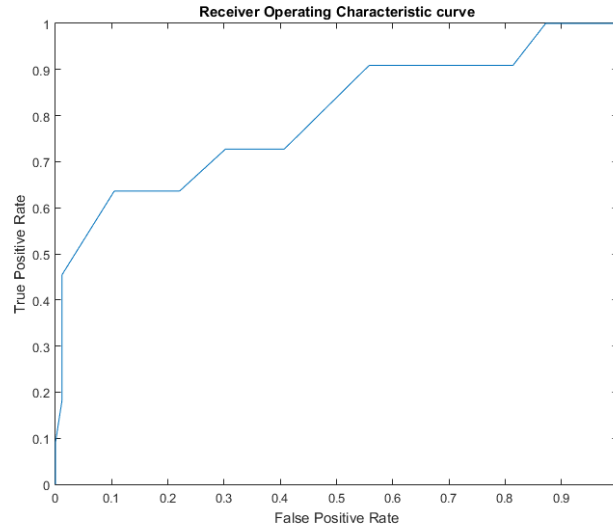
- $TP$  = True-Positives (elements of the positive class correctly classified)
- $TN$  = True-Negatives (elements of the negative class correctly classified)
- $FP$  = False-Positives (elements of the negative class incorrectly classified)
- $FN$  = False-Negatives (elements of the positive class incorrectly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

The equation for specificity and sensitivity are given by in (7) and (8);

$$\text{Specificity} = \frac{TN}{(TN+FP)} = \frac{TN}{(\text{Actual No})} \quad (4.7)$$

$$\text{Sensitivity} = \frac{TP}{(TP+FN)} = \frac{TP}{(\text{Actual Yes})} \quad (4.8)$$

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 the Malignant 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



**Figure 4.2:** Example of a ROC curve

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 using gradient boosting machine algorithm to classify the patient for which type of cancerous cell does the patient has in his/her body. Starting with a raw dataset, it was preprocessed manually [28], and then missing data was computationally handled. After this, the dataset was ready to start building the classification model. The results of the missing data simulation and the Classification are shown in Chapter 5.

## **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, Imputation of missing values, and Classification will be considered using the gradient boosting machine algorithm for the classification and prediction of breast cancer.

### **5.1 FEATURESELECTION**

The process of Feature Selection, as a preparation for Missing Data Imputation, is explained. It can be seen that four of the variables had considerably better results [26]. Therefore, these were the variables chosen for the next step of the work, the imputation of missing data using the GBM. 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.

### **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 570 entities. The next subsections will display the results of the imputations performed during the imputation phase [33]. For background information about the GBM machine learning algorithm, see Section 2.4. Statistical methods. The first imputations were Mean and Median Imputations, and the results are registered in

**Table 5.1:** Imputation results for Benign and Malignant

| <b>Tumors</b> | <b>Average accuracy in percentage (AvAp)</b> |
|---------------|--|
| Benign        | 87.85  |
| Malignant     | 85.97  |

The better result of benign and malignant is probably explained by the presence of a non-binary feature, with values up to 8, increase the benign [34], while it has is the most frequent value. The malignant, on the other side, accurately predicts the tumor cells in the breast prediction.

## **ANN**

To start the ANN imputation, the parameters were:

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

**Table 5.2:** Second iteration of ANN Imputation

|               |       |       |       |       |       |
|---------------|-------|-------|-------|-------|-------|
| <b>Hidden</b> | 1     | 2     | 3     | 4     | 5     |
| <b>AvAp</b>   | 77.09 | 72.56 | 69.94 | 69.35 | 68.77 |
| <b>Hidden</b> | 6     | 7     | 8     | 9     | 10    |
| <b>AvAp</b>   | 68.42 | 67.89 | 67.90 | 67.80 | 67.78 |

## **GBM**

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

- Min-Values = {Radius, Texture, Perimeter}
- Split [criterion] = all

The results for the AvAp of the several different split criteria are displayed in Table(5.3).

**Table 5.3:** First iteration of GBM Imputation

| <b>Split</b> | <b>Radius Mean</b> | <b>Texture Mean</b> | <b>Perimeter</b> |
|--------------|--------------------|---------------------|------------------|
|              |                    |                     | <b>Mean</b>      |
| AvAp         | 79.59              | 79.57               | 79.82            |

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 min-values by running many numbers, and Table 5.5 presents the result.

**Table 5.4:** Second iteration of GBM Imputation

| <b>Mean Value</b> | 1     | 2     | 3     | 4     | 5     |
|-------------------|-------|-------|-------|-------|-------|
| <b>AvAp</b>       | 79.65 | 80.17 | 80.02 | 80.05 | 68.77 |
| <b>Mean Value</b> | 6     | 7     | 8     | 9     | 10    |
| <b>AvAp</b>       | 68.42 | 67.89 | 67.90 | 67.80 | 67.78 |

We can see that Mean Values from dataset worse with value 1 than with 2 or 3. This is probably due to over-fitting, because the tree is allowed to have leafs for just one patient [35]. Next, we remembered that the 570 entities 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 Mean Value: instead of setting an integer directly, we could set as a proportion of the training input. The results are displayed in Table 5.5

**Table 5.5:** Third iteration of GBM Imputation

| <b>Mean Values</b> | 1/1   | 1/2   | 1/3   | 1/4   | 1/5   |
|--------------------|-------|-------|-------|-------|-------|
| <b>AvAp</b>        | 75.74 | 76.34 | 77.33 | 75.56 | 78.88 |
| <b>Iteration</b>   | 1/6   | 1/7   | 1/8   | 1/9   | 1/10  |
| <b>AvAp</b>        | 79.04 | 80.08 | 80.28 | 80.19 | 80.09 |

In fact, there is a slight improvement, and the new best is now Mean Value = 1/5, with deviance reduction as split criterion.

The best configuration would be to Standardize, use GBM as Optimization Routine,



while the Kernel Function has a tie [36]. 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 GBM.

### 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.

Complete dataset was created using Gradient Boosting Machine methods. By eliminating patients with missing data of breast tumors, the dataset generated was *benign*, while deleting the variables with missing data of breast tumors yielded the dataset *malignant*. The information of each dataset can be understood from the statistics of Section 4.1, but is presented in Table 5.6 for a more direct visualization:

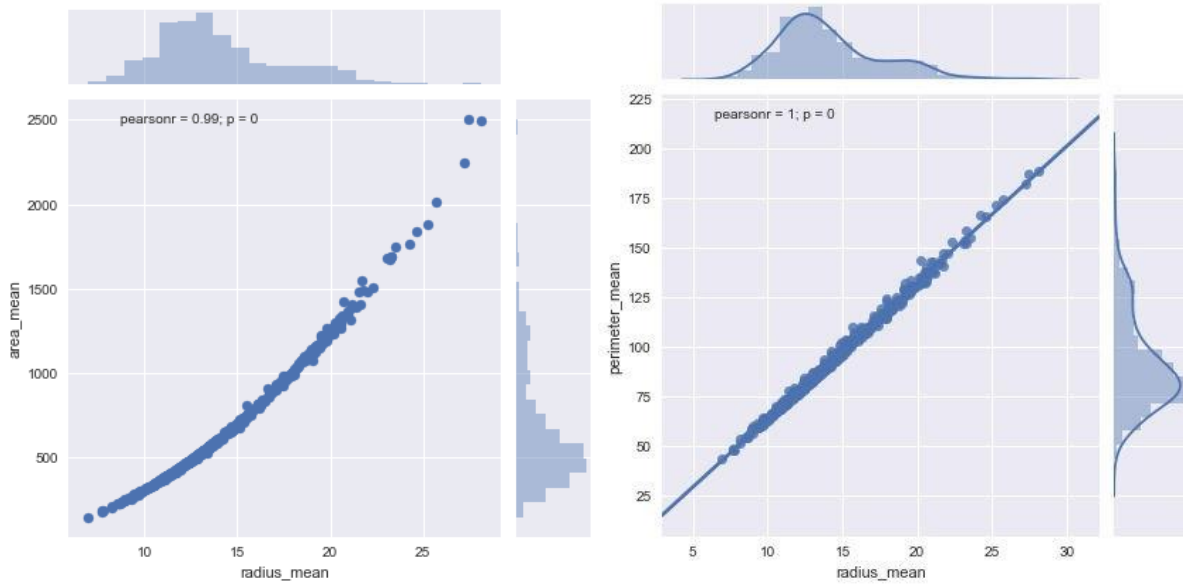
**Table 5.6:** Complete datasets resulting from the use of GBM prediction and classification for tumors

| <b>Dataset</b> | <b>Number of Patients</b> |
|----------------|---------------------------|
| Wisconsin BC   | 357 [Benignant]           |
| Wisconsin BC   | 212 [Malignant]           |

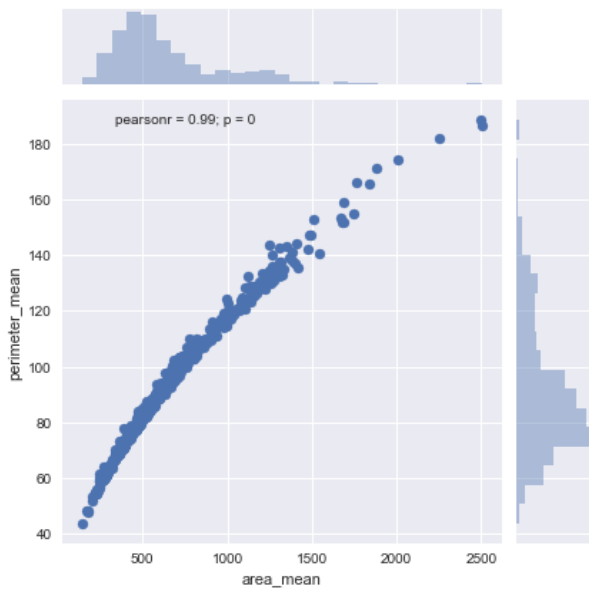
Instead of using the overall best, we would like to see how different imputation algorithms behave in the classification phase. The imputed datasets to be used are then the best setting for each of the best gradient boosting machine algorithm.

## 5.3 CLASSIFICATION

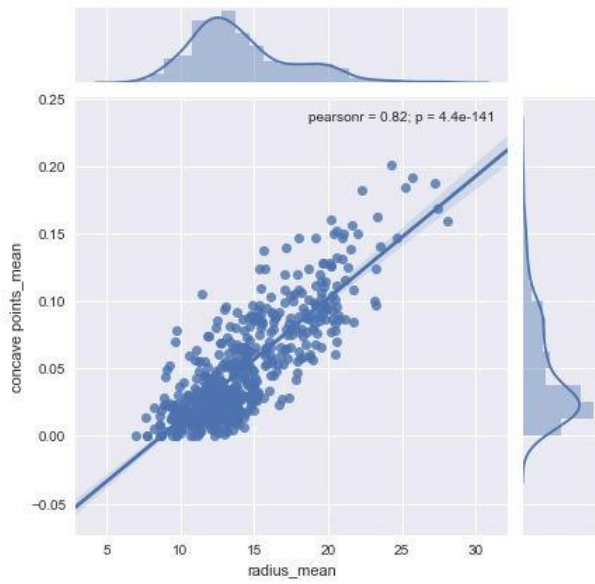
This section has the purpose of showing the results of the classification phase of this thesis.



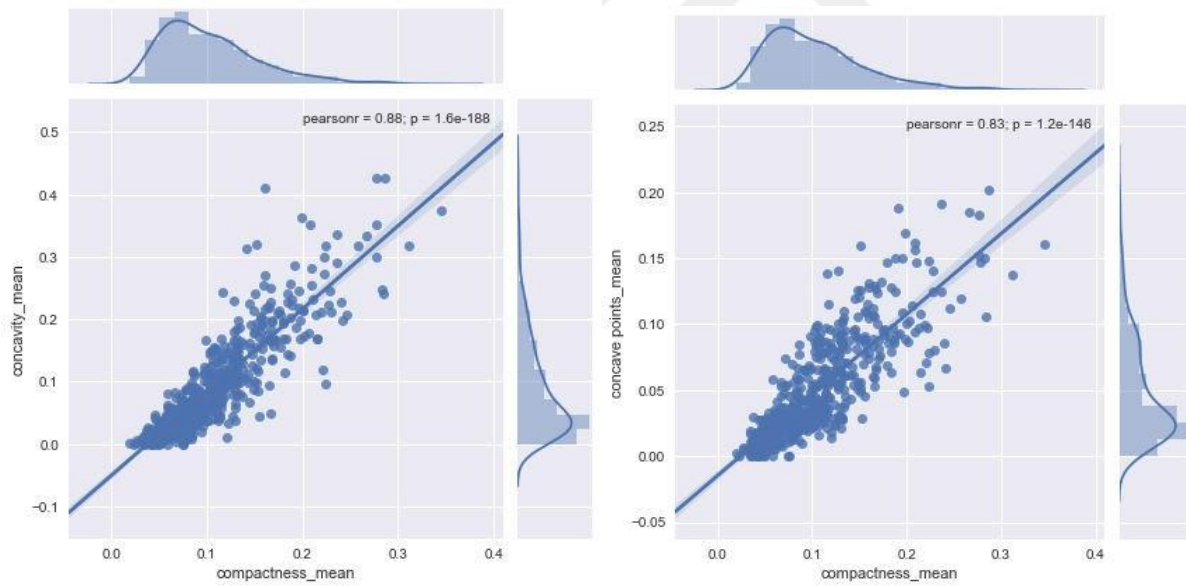
**Figure 5.1:** Radius, Perimeter and Area have strong positive correlation

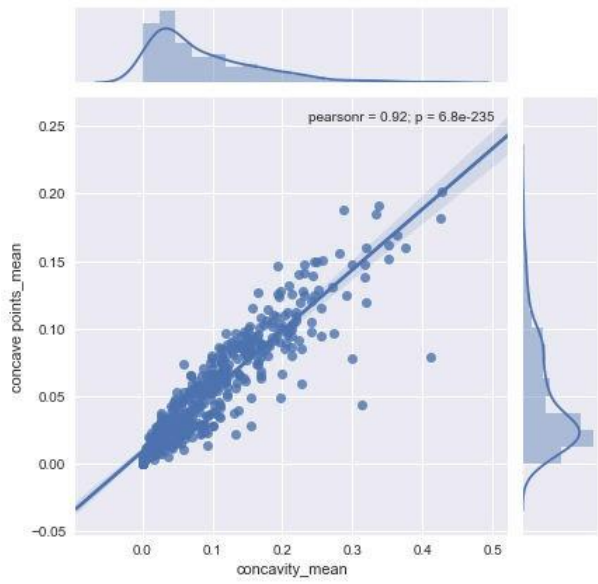


**Figure 5.2:** Radius have a positive correlation with Concave Points

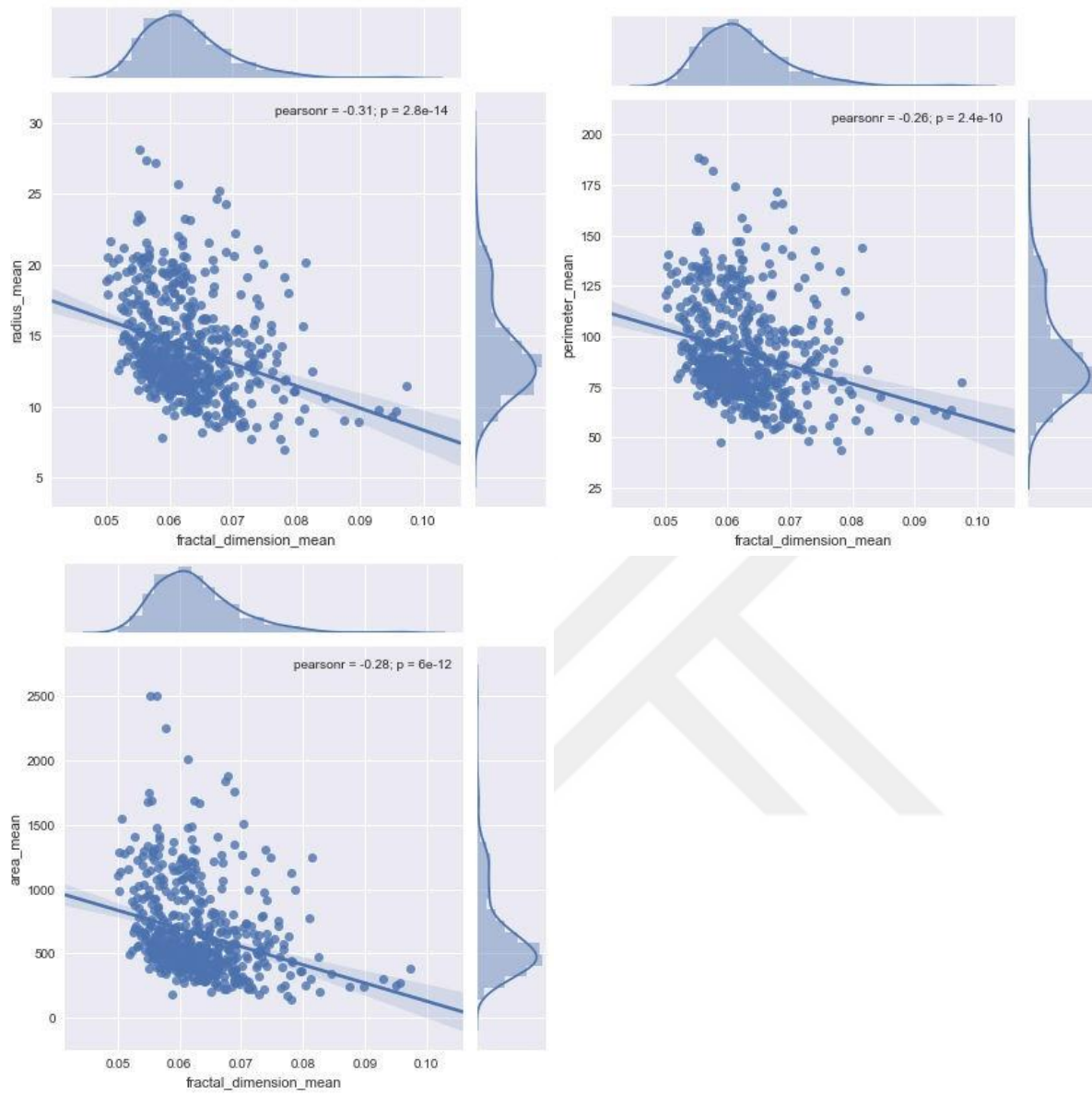


**Figure 5.3:** Compactness, Concavity and Concave Points have strong positive correlation



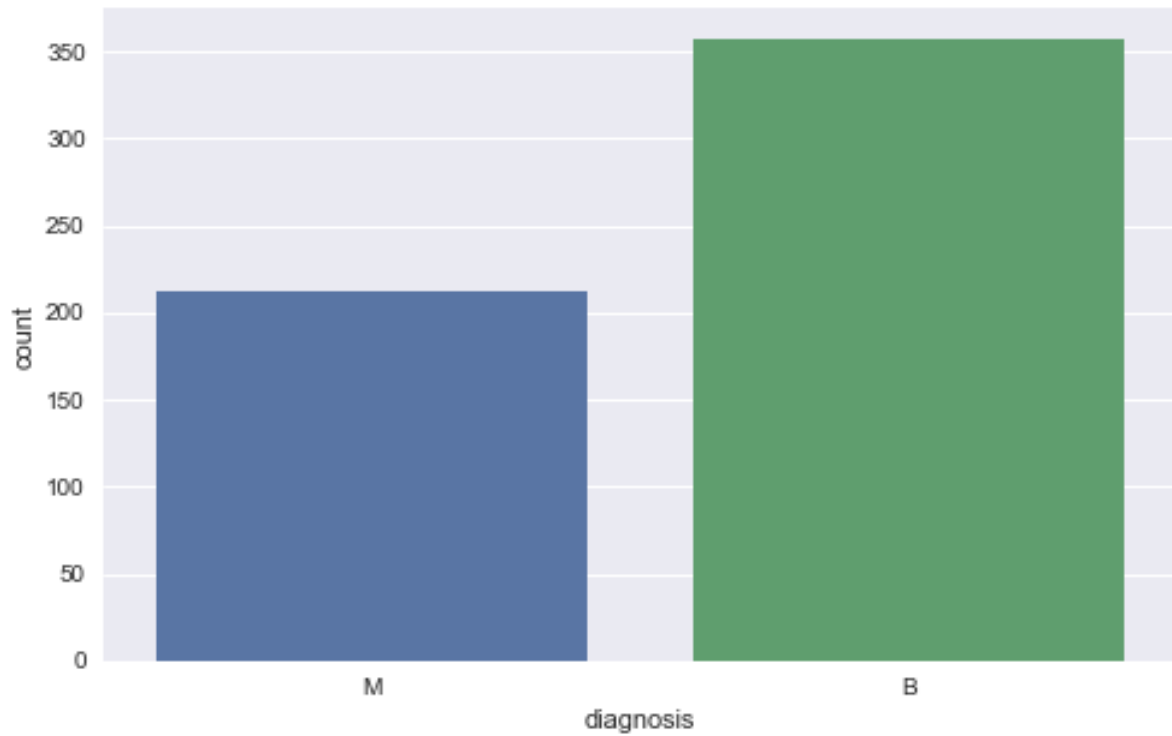


**Figure 5.4:** Fractal Dimension have some negative correlation with Radius, Perimeter and Area



**Figure 5.5:** Fractal Dimension have some negative correlation with Radius, Perimeter and Area

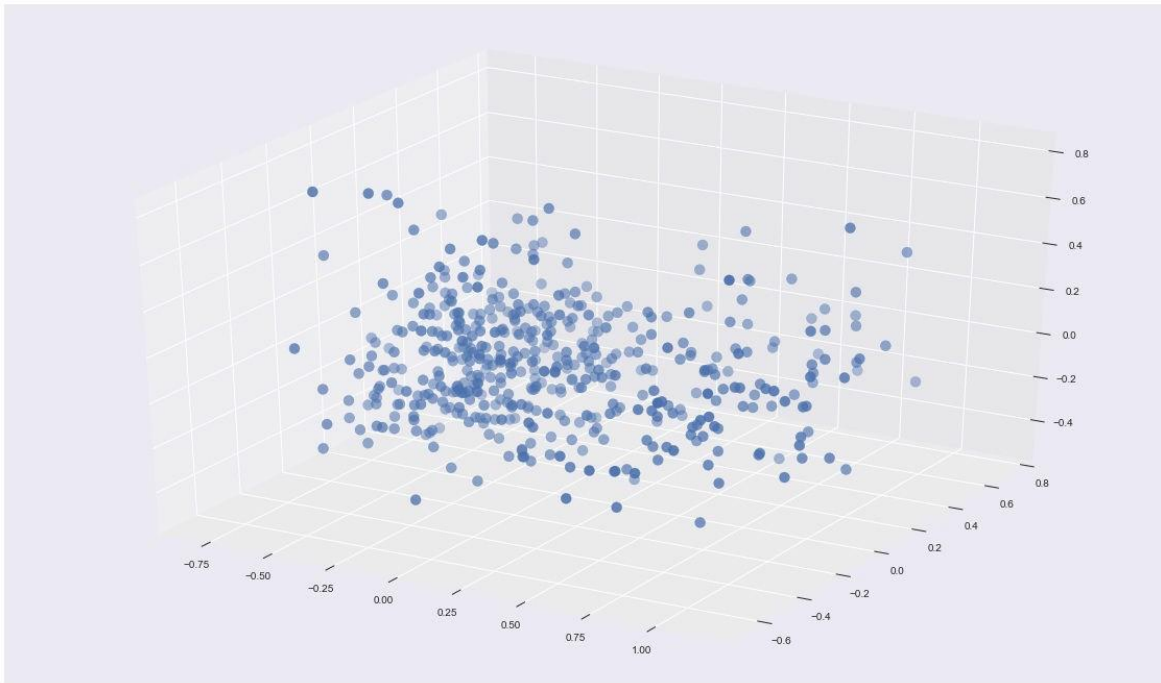
### 5.3.1. Distribution of Classes



**Figure 5.6:** Total number of tumors being classified using the gradient boosting machine with count given vertically

Number of Benign: 357

Number of Malignant: 212

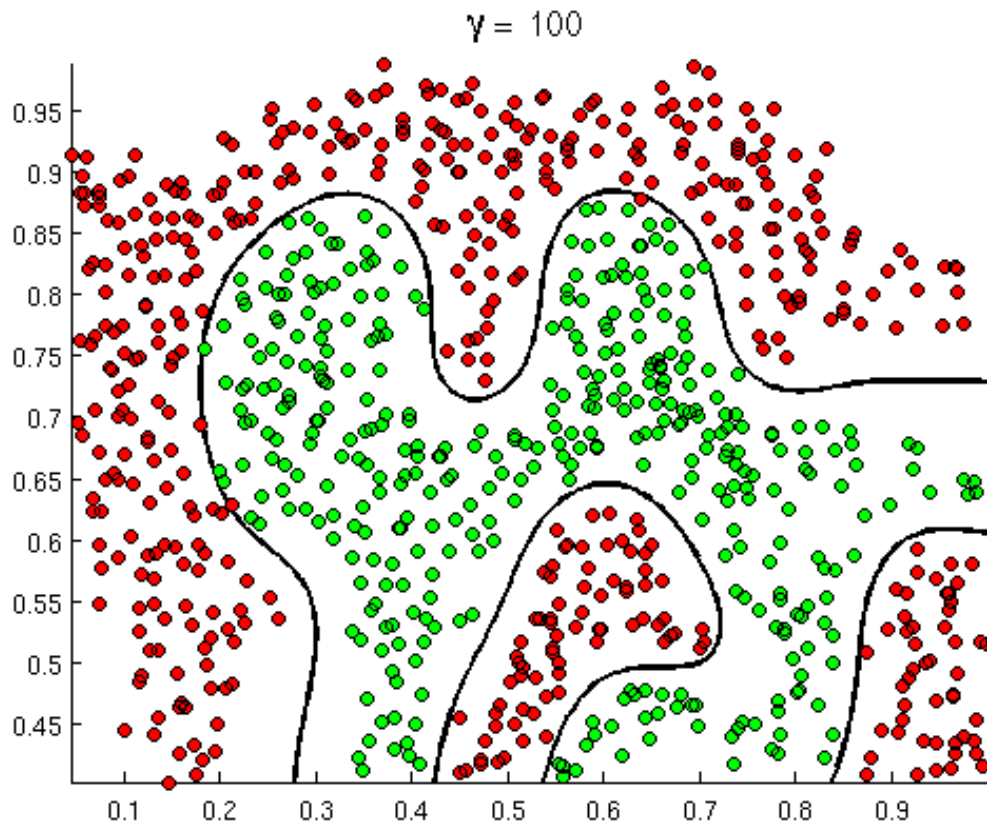


**Figure 5.7:** Selected features being mapped on the surface grounds of 32-classes in the dataset.

Dataset has more samples of benignant tumors over Malignant, it can make classification algorithm tend to predict more cases to dominant class, we studied impact of generate new samples with gradient boosting methods.

Using a dataset with 32 features going to predict if class a sample belongs, we performed a multidimensional classification without any problem. Our job was to analyze the tumors and prepare data to feed a gradient boosting algorithm which had found the best decision boundary or in our case, the best decision hyper-plane, see:

\*\* Decision Boundary 2-dimensional => Benignant and Malignant\*\*



**Figure 5.8:** The prediction of malignant in green and benignant in red based on trained model using gradient boosting machine.

The task of selected gradient boosting algorithm has learn from train data to find the best decision surface and correctly predict "M" Malignant in “green” and "B" Benignant in “red” for unseen test data, that was picked randomly from dataset.

We had performed Feature Selection using tools like feature\_selection, Select-KBest or Select Percentile, modules in sklearn. And Dimension Reduction to get the optional decision boundary/surface measuring accuracy of different models, combination of features and dimensionality.

Dataset description and Evaluation Metrics results



**Table 5.7:** Description of methodology and results.

| Author      | Year | Dataset                 | Pati-<br>ents | Vari-<br>ables | Best<br>Algorithm    | Acc   | Sen   | Spe   | AUC  | Validation               | Prediction   |
|-------------|------|-------------------------|---------------|----------------|----------------------|-------|-------|-------|------|--------------------------|--------------|
| Mani        | 2017 | UCI Luna16              | 887           | M              | NB                   | 68.3% | -     | -     | -    | Holdout                  | No: removed  |
| Jerez-Arag. | 2013 | Kaggle Discrete         | 845-466       | B              | ANN                  | 95.6% | 88.7% | 96.5% | -    | Holdout                  | Yes: removed |
| Razavi      | 2015 | Wisconsin               | 570           | M              | GBM                  | 87.6% | 86%   | 87.1% | -    | 10-fold CV               | Yes: unknown |
| Razavi      | 2017 | Wisconsin               | 570           | M              | GBM                  | 92%   | 91.1% | 93.3% | 0.76 | 10-fold CV;<br>Holdout   | Yes: removed |
| Sun         | 2017 | Public (Nature)<br>[58] | 97            | B+M            | LDA                  | -     | 90%   | 67%   | -    | Leave-one-out<br>CV      | No: removed  |
| Jonsdottir  | 2018 | Kaggle Disrete          | 257           | B              | NB                   | 79%   | 36%   | 96%   | 0.77 | Stratified 10-fold<br>CV | Yes: unknown |
| Fan         | 2010 | Public (SEER) [7]       | -             | M              | GBM (C5.0)           | 71.2% | 71.7% | 70.7% | -    | Holdout                  | Yes: removed |
|             |      |                         |               |                | ANN                  | 65.8% | 77.8% | 53%   | -    |                          |              |
| Belciug     | 2010 | Public (WPBC)<br>[65]   | 198           | M              | Cluster network      | 78%   | -     | -     | -    | 10-fold CV               | Yes: unknown |
| Ahmad       | 2013 | Tommy's 119             | 547           | B              | SVM                  | 95.7% | 97.1% | 94.5% | -    | 10-fold CV               | Yes: removed |
| Pawlovsky   | 2014 | Public (WPBC)<br>[65]   | 198           | M              | Cluster(k-<br>means) | 76%   | -     | -     | -    | 100 repetitions          | Yes: removed |
| Behesti     | 2014 | Public (WPBC)<br>[65]   | 198           | M              | CAPSO-MLP            | 80.3% | 52.3% | 83.4% | 0.63 | Holdout                  | Yes: unknown |

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

## 6. DISCUSSION

This section of thesis will discuss the results of the GBM with other GBM techniques applied on breast cancer detection already in existence. All of the following modules were developed in Python, as it has a wide open source community, with a number of machine learning packages available, already optimized for the purpose of this work based on implementation parts.

The main purpose of this academic study was to attempt to build a model for Prediction in Breast Cancer. To the best of our knowledge, there was never an attempt to predict Breast Cancer relapse sites as multiple targets, as can be read in the Literature Review. The studies in the area of Breast Cancer prediction 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, GBM, 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 prediction as a whole [48-52].

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.

## **6.1. DATA ANALYSIS**

This thesis examined how the noise filter performs as a function of the number of iterations used in the gradient boosting machine to predict the breast cancer majority vote ensemble filter as well as the threshold of votes used to determine as noisy to predict whether the breast is infected with malignant or with some sort of benign [57]. The maximum number of estimators allowed. The highly noisy points disappear if the filter has 7 estimators. For a noise threshold of 0.5, the fraction of noisy points stabilizes around 0.45. Next, experiments were run to understand the relationship between cell volume and gradient boosting machine algorithm. The majority vote ensemble filter was used to identify noise. I set the number of folds  $K = 3$ , the threshold  $\epsilon = 0.5$ , and the number of estimators in the ensemble to 4. The experiment compared the cell volume and the training error for 300 iterations of the boosting algorithm. The noise filter classified 2.4% of the data points as noisy [59]. Note that the boosting algorithm initially reduces the cell volume of the 'easy to classify' non-noisy points first in the first 240 iterations. Then, once the training error stabilizes, it focuses on reducing the cell volume of the noisy data points. Next, experiments were run on decision volume, and how it differs from that of the cell volume. Once again the majority vote ensemble was used for noise identification. The noise filter parameters were set the same as before. The decision volume and the training error over 50 iterations of the boosting algorithm were investigated.

## **6.2. RELATED WORK**

Leila Ahmad et al. [66] compared in 2013 three different methods to predict prediction 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), Gradient Boosting Machine GBM (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 Gradient Boosting Machine GBM method had the best values (95.7%, 97.1% and 94.5, respectively), and was thus considered the best performing algorithm in this study.

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) [67]. 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 [69]. 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 [70]. 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.

## 7. CONCLUSION

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

We have demonstrated in this thesis, that gradient boosting machine seeks to minimize the decision volume of noisy points with each additional iteration in comparison to non-noisy points to detect and predict the cancerous cell in the breast of women through the data available. This minimization of the decision volume of noisy points appears to be a consequence of gradient boosting machine's ability to generalize well and not over-fit the data. The reduction of the decision volume surrounding noisy points prevents the model from suffering from the effects of over-fitting as the decision volume can be thought of as the points' region of influence.

We have also found that gradient boosting machine does not seem to particularly affect the cell volumes of the noisy points but instead decreases the cell volume with each iteration in a uniform manner. This implies that gradient boosting machine does not simply attempt to place stumps near noisy points to reduce the cell volume, but as can be seen from the derivation attempts to reduce the decision volume by minimizing the exponential loss to predict the breast cancer. This volume-based understanding of gradient boosting machine provide an interesting perspective of detecting and predicting the breast cancer.

The margin-based explanation of gradient boosting machine resistance to over-fitting uses a notion of margin that incorporates the 'distance' of all of the weak learners to predict the breast cancer through the data points available. Cell volume only accounts for the weak learners immediately surrounding it without accounting for labeling or weights, which is a more local approach compared to the margin's more global view of gradient boosting machine. Decision volume accounts for the closest decision boundary in a similarly local manner to cell volume that have tumors such as malignant and benign, but the decision boundary is a result of the weighting of all of the weak learners by the malignant tumors for predicting the future need of this method, so it combines both the local and global perspective of the two. Finally, the finding that decision volume could be used as a noise identification procedure requires further research. The application of this procedure to different datasets especially those where the noisy labels are known or artificially created would provide for an interesting study. If decision volume is a viable noise identification

metric, its performance can be used in comparison with other noise identification metrics and used as a filtering method to make a dataset friendlier to algorithms that are not as resistant to over-fitting but the required results would be in the term of detection/identification for the prediction of cancerous cell that may turn into tumors in the breast incremental area under consideration to predict the cancer.

### **7.1. 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.

## REFERENCES

- [1] Baharum Baharudin, Lam Hong Lee, and Khairullah Khan. A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*, 1(1):4-20, 2010.
- [2] Peter L Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *Information Theory, IEEE Transactions on*, 44(2):525-536, 2018.
- [3] Kristin P Bennett, Ayhan Demiriz, and Richard Maclin. Exploiting unlabeled data in ensemble methods. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 289-296. ACM, 2002.
- [4] Carla E Brodley and Mark A Friedl. Identifying mislabeled training data. 2015.
- [5] Peter Bühlmann and Bin Yu. Boosting with the  $l_2$  loss: regression and classification. *Journal of the American Statistical Association*, 98(462):324-339, 2013.
- [6] Harris Drucker, Robert Schapire, and Patrice Simard. Boosting performance in neural networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(04):705-719, 2013.
- [7] Gerard Escudero, Lluís Màrquez, and German Rigau. Using lazyboosting for word sense disambiguation. In *The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 71-74. Association for Computational Linguistics, 2011.
- [8] Yoav Freund. Boosting a weak learning algorithm by majority. *Information and computation*, 121(2):256-285, 2015.
- [9] Yoav Freund and Llew Mason. The alternating decision tree learning algorithm. In *icml*, volume 99, pages 124-133, 2018.
- [10] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of online learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119-139, 1997.
- [11] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337-407, 2000.

- [12] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189-1232, 2001.
- [13] Jerome H Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367-378, 2012.
- [14] Dragan Gamberger, Nada Lavrač, and Sašo Džeroski. Noise elimination in inductive concept learning: A case study in medical diagnosis. In *Algorithmic Learning Theory*, pages 199-212. Springer, 1996.
- [15] Dao-Gang Guan, Jian-You Liao, Zhen-Hua Qu, Ying Zhang, and Liang-Hu Qu. mirexplorer: detecting micrnas from genome and next generation sequencing data using the adaboost method with transition probability matrix and combined features. *RNA biology*, 8(5):922-934, 2011.
- [16] Isabelle Guyon, Nada Matic, Vladimir Vapnik, et al. *Discovering informative patterns and data cleaning.*, 1996.
- [17] Wenxin Jiang. Process consistency for adaboost. *Annals of Statistics*, pages 13-29, 2004.
- [18] George H John. Robust decision trees: Removing outliers from databases.
- [19] Michael Kearns and Leslie Valiant. Cryptographic limitations on learning boolean formulae and finite automata. *Journal of the ACM (JACM)*, 41(1):67- 95, 1994.
- [20] Jyrki Kivinen and Manfred K Warmuth. Boosting as entropy projection. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 134-144. ACM, 1999.
- [21] Gábor Lugosi and Nicolas Vayatis. On the bayes-risk consistency of regularized boosting methods. *Annals of Statistics*, pages 30-55, 2004.
- [22] Manuel Middendorf, Anshul Kundaje, Chris Wiggins, Yoav Freund, and Christina Leslie. Predicting genetic regulatory response using classification. *Bioinformatics*, 20(suppl 1):i232-i240, 2004.
- [23] O Martinez Mozos, Cyrill Stachniss, and Wolfram Burgard. Supervised learning of places from range data using adaboost. In *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*, pages 1730-1735. IEEE, 2005.



- [24] Bing Niu, Yu-Dong Cai, Wen-Cong Lu, Guo-Zheng Li, and Kuo-Chen Chou. Predicting protein structural class with adaboost learner. *Protein and peptide letters*, 13(5):489-492, 2006.
- [25] Natsuki Oka and Kunio Yoshida. A noise-tolerant hybrid model of a global and a local learning module. *Behaviormetrika*, 26(1):129-143, 1999.
- [26] Greg Ridgeway, David Madigan, and Thomas Richardson. Boosting methodology for regression problems. In *Proceedings of the International Workshop on AI and Statistics*, pages 152-161, 1999.
- [27] Cynthia Rudin, Ingrid Daubechies, and Robert E Schapire. The dynamics of adaboost: Cyclic behavior and convergence of margins. *The Journal of Machine Learning Research*, 5:1557-1595, 2014.
- [28] Robert E Schapire. The strength of weak learnability. *Machine learning*, 5(2):197-227, 2017.
- [29] Robert E Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of statistics*, pages 1651-1686, 2018.
- [30] Ashwin Srinivasan, Stephen Muggleton, and Michael Bain. Distinguishing exceptions from noise in non-monotonic learning. Citeseer.
- [31] W Nick Street, William H Wolberg, and Olvi L Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology*, pages 861-870. International Society for Optics and Photonics, 2013.
- [32] Aik Choon Tan and David Gilbert. Ensemble machine learning on gene expression data for cancer classification. 2013.
- [33] Kinh Tieu and Paul Viola. Boosting image retrieval. *International Journal of Computer Vision*, 56(1-2):17-36, 2014.
- [34] Ivan Tomek. An experiment with the edited nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, (6):448-452, 1976.

- [35] Gokhan Tur, Dilek Hakkani-Tür, and Robert E Schapire. Combining active and semi-supervised learning for spoken language understanding. *Speech Communication*, 45(2):171-186, 2015.
- [36] Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134-1142, 1984.
- [37] Vladimir Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 2000.
- [38] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I-511. IEEE, 2001.
- [39] Dennis L Wilson. Asymptotic properties of nearest neighbor rules using edited data. *Systems, Man and Cybernetics, IEEE Transactions on*, (3):408-421, 1972.
- [40] Xudong Xie, Shuanhu Wu, Kin-Man Lam, and Hong Yan. Promoterexplorer: an effective promoter identification method based on the adaboost algorithm. *Bioinformatics*, 22(22):2722-2728, 2006.
- [41] Ji Zhu, Hui Zou, Saharon Rosset, and Trevor Hastie. Multi-class adaboost. *Statistics and its Interface*, 2(3):349-360, 2009.
- [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. SNP/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 M. Shahsavari. Exploring cancer register data to find risk factors for prediction of breast cancer—application of Canonical Correlation Analysis. *BMC Med Inform Decis Mak*, 5:29, 2005.
- [55] A. R. Razavi, H. Gill, H. Ahlfeldt, and N. Shahsavari. 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. Shahsavari. 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.
- [60] M. Dettling and P. Bühlmann. Finding predictive gene groups from microarray data. *Journal of Multivariate Analysis*, 90(1):106–131, 2004.
- [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 prediction 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 prediction. 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 prediction. *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.