**Universidade do Minho**
Escola de Engenharia

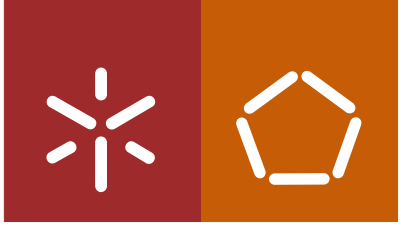
Ana Paula Pinto da Silva

# A Survival Prediction Model for Colorectal Cancer Patients

July 2016

Ana Paula Pinto da Silva **A Survival Prediction Model forColorectal Cancer Patients**

UMinho | 2016



**Universidade do Minho**
Escola de Engenharia

Ana Paula Pinto da Silva

# A Survival Prediction Model for Colorectal Cancer Patients

Master Dissertation
Integrated Master in Biomedical Engineering

Dissertation supervised by
**Paulo Jorge Freitas de Oliveira Novais**

July 2016

DECLARAÇÃO

Nome

Ana Paula Pinto da Silva

Endereço electrónico: silva.anapp@gmail.com

Número do Bilhete de Identidade: 13742458

Título dissertação :

A Survival Prediction Model for Colorectal Cancer Patients

Orientador :

Paulo Jorge Freitas de Oliveira Novais

Ano de conclusão: 2016

Designação do Mestrado ou do Ramo de Conhecimento do Doutoramento:

Mestrado Integrado em Engenharia Biomédica

1. É AUTORIZADA A REPRODUÇÃO INTEGRAL DESTA DISSERTAÇÃO APENAS PARA EFEITOS DE INVESTIGAÇÃO, MEDIANTE DECLARAÇÃO ESCRITA DO INTERESSADO, QUE A TAL SE COMPROMETE;

Universidade do Minho, 11 / 07 / 2016

Assinatura: Ana Silva

# AGRADECIMENTOS

Agradeço aos meus orientadores. Ao Professor Paulo Novais pelo profissionalismo e perseverança e, ao Tiago Oliveira por todo o acompanhamento, rigor e sobretudo pela sua dedicação.

Agradeço ainda ao Dr. Pedro Leão, o seu contributo foi fundamental.

Aos que fazem parte do ISLab, pela disponibilidade e partilha de conhecimento que foi parte integrante desta tese.

Aos que me são próximos e acreditaram, mesmo quando eu duvidei.

À minha família. Pai, mãe e irmã, por tudo. Espero um dia poder retribuir minimamente tudo o que fizeram por mim e, um dia ser um exemplo para a minha sobrinha e afilhada.

Dedico a dissertação aos meus pais.

## ABSTRACT

The importance of making predictions in health is mainly linked to the decision-making process. Make survival predictions accurately is a very difficult task for healthcare professionals and a major concern for patients. On the one hand, it can help physicians decide between palliative care or other medical practice for a patient. On the other hand, the notion of remaining lifetime could help patients in the realization of dreams. However, the prediction of survivability is directly related to the experience of health professionals and their ability to memorize.

Most decisions are made based on probability and statistics, but these are based on large groups of people and may not be suitable to predict what will happen in particular cases. Consequently, the use of *machine learning* techniques have been explored in healthcare. Their ability to help solve diagnostic and prognosis problems has been increasingly exploited.

The main contribution of this work is a prediction tool of survival of patients with cancer of the colon and/or rectum, after treatment and a few years after treatment. The characteristics that distinguishes it is the balance between the number of required inputs and their performance in terms of prediction. The tool is compatible with mobile devices, includes a online learning component that allows for automatic recalculation and flexibly of the prediction models, by adding new cases.

The tool aims to facilitate the access of healthcare professionals for instruments that enrich their practice and improve their results. This increases the productivity of healthcare professionals, enabling them to make decisions faster and with a lower error rate.

# RESUMO

A importância de fazer previsões na área da saúde está sobretudo ligada ao processo de tomada de decisão. Fazer previsões de sobrevivência de forma precisa é uma tarefa muito difícil para os profissionais de saúde e uma grande preocupação para os pacientes. Por um lado, pode ajudar os médicos a decidir entre cuidados paliativos ou outra prática médica para um paciente. Por outro lado, a noção do tempo de vida remanescente poderia ajudar os pacientes na concretização de sonhos. No entanto, este tipo de previsão está diretamente relacionada com a experiência do profissional de saúde e da sua capacidade de memorizar.

A maior parte das decisões são tomadas com base em probabilidades e estatística, mas estas têm como base grandes grupos de pessoas, podendo não ser adequadas para prever o que vai acontecer em casos particulares. Por conseguinte, a utilização de técnicas de *machine learning* têm sido exploradas na área da saúde. A sua capacidade para ajudar a resolver problemas de diagnóstico e prognóstico tem sido cada vez mais explorada.

A principal contribuição deste trabalho é uma ferramenta de previsão da sobrevida de pacientes com cancro do cólon e/ou do reto, após o tratamento e alguns anos após o tratamento. As características que a distingue são o equilíbrio entre o número de entradas necessárias e o seu desempenho a nível da previsão. A ferramenta, compatível com dispositivos móveis, possui uma componente de aprendizagem em tempo real que permite recalcular de forma automática e evolutiva os modelos usados para fazer a previsão, através da adição de novos casos.

A ferramenta tem como propósito facilitar o acesso dos profissionais de saúde a instrumentos capazes de enriquecer a sua prática e melhorar os seus resultados. Esta aumenta a produtividade dos profissionais de saúde, permitindo que estes tomem decisões mais rapidamente e com uma taxa de erro menor.

# CONTENTS

# LIST OF FIGURES

**List of Figures**

# LIST OF TABLES

# LIST OF EQUATIONS

# ACRONYMS

**A**

**ACCENT** Adjuvant Colon Cancer End Points.

**AI** Artificial Intelligence.

**AIM** Artificial Intelligence in Medicine.

**AJCC** American Joint Committee on Cancer.

**ANN** Artificial Neural Network.

**API** Application Programming Interface.

**AUC** The Area Under the ROC Curve.

**C**

**C-INDEX** Concordance Index.

**CC** Colon Cancer.

**CFS** Correlation Feature Selection.

**CRC** Colorectal Cancer.

**I**

**IGR** Information Gain Ratio.

**M**

**ML** Machine Learning.

**ML-BBN** Machine-Learned Bayesian Belief Network.

**N**

**NAP** Nodes + Prognostic Factors.

**NCDB** National Cancer Data Base.

**Acronyms**

**R**

**RC** Rectal Cancer.

**ROC** Receiver Operating Characteristic.

**S**

**SEER** Surveillance, Epidemiology, and End Results.

**SMOTE** Synthetic Minority Over-sampling Technique.

# 1

# INTRODUCTION

Health care professionals are confronted daily with new diseases, new therapeutics, quick decisions and cost reductions. At the same time, technology has an increasingly important role in answering these challenges by finding new solutions, supporting health care professionals in the course of their duties, assisting with tasks of data and knowledge manipulation, testing new treatments, simulating scenarios, and developing new devices.

This dissertation discloses and assistive tool to help physicians improve their practice. The problem it addresses is predicting the survival of *Colorectal Cancer (CRC)* patients (in an individualized manner). The knowledge upon which the features of the tool are based was drawn from a large volume of collected data from patients.

This chapter provides a deeper understanding about CRC and the motivation, objectives, methodology and research behind this dissertation.

## 1.1 PRELIMINARY NOTIONS

### 1.1.1 *Anatomy and Physiology of the Digestive System*

The cell is the basic unit of life. It is the smallest unit capable of all of the processes that define life. A cell grows and divides into new cells, in order to ensure the proper body functions, replacing worn out or injured cells. If a damaged cell is not repaired or does not die, it continues to grow and forms new abnormal cells. When a cell grows out of control and invades other tissues it is called a cancer cell. These cells can break away and travel to other parts of the body through the bloodstream or the lymph system, growing and forming new tissues there. This spread of cancer to new areas of the body is called metastasis.

The digestive system (Figure 1) consists of the digestive tract and its accessory organs (such as the teeth), where food is processed into molecules that can be absorbed to give the body cells the energy and other substances they need to operate. The digestive tract, also known as the gastrointestinal tract, consists of a long continuous tube that begins at

the mouth, includes the pharynx, esophagus, stomach, small intestine, large intestine, and ends at the anus [7, 87].

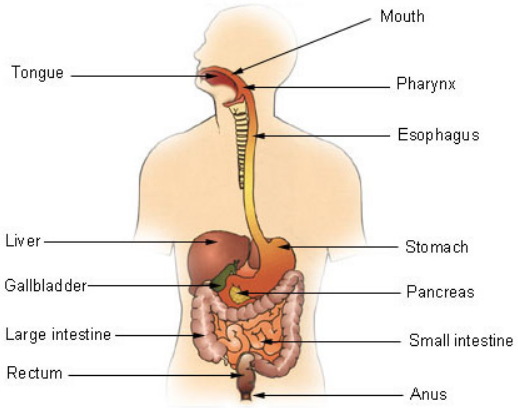



Figure 1.: Organs of the digestive system [95].

The colon is the most extensive part of the large intestine, also known as bowel. Its function is to absorb water and nutrients from the food matter and to serve as a storage place for waste matter. As shown in Figure 2, the colon has 4 parts. The first section is the ascending colon. It starts with a small pouch (the cecum), where the small intestine joins the colon, and it extends upward on the right side of the abdomen. The second part is called the transverse colon, as it goes across the body from the right to the left side in the upper abdomen. The third section, called the descending colon, continues downward on the left side. Finally, the fourth and last part is known as the sigmoid colon because for its ”S” or ”sigmoid” shape. After going through the colon, the material that is left is called feces or stool. The material enters the rectum, where it is stored until it is expelled of the body through the anus [6].

The wall of the colon and rectum (Figure 3) is made up of several layers: the mucosa (or mucous membrane), the submucosa,the muscularis (or muscularis propria), and the adventitia (or serosa). The mucosa is the innermost layer of the digestive tract that is in direct contact with digested food. Its layers are responsible for most of digestion, absorptive and secretory processes, and also for passing waste matter. The submucosa consists in a dense irregular layer of connective tissue that contains fibroblasts, mast cells, blood and lymphatic vessels, and a nerve fiber plexus. The muscular layer is the main responsible for the contractility. It consists of an inner circular layer that prevents food from traveling backward, and a longitudinal outer layer that shortens the tract. Finally, the serosa forms the outermost layer of the gastrointestinal tract. It consists of several layers of connective tissue and secretes a fluid in order to lubricate the surface of the large intestine, defending

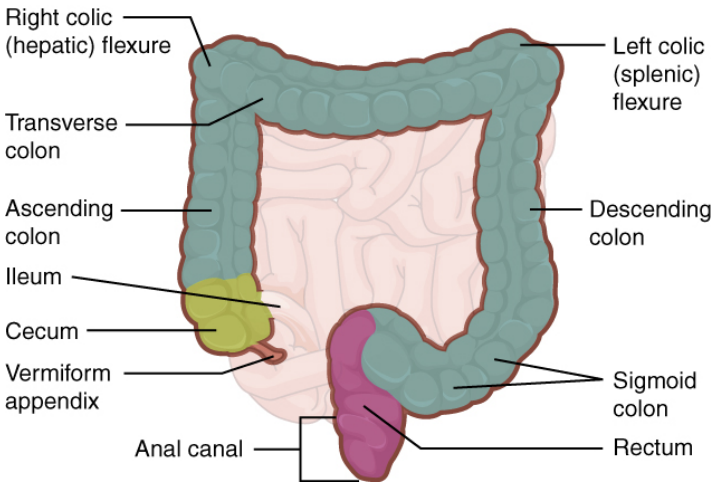



Figure 2.: The large intestine [6].

it from the friction between abdominal organs and the surrounding muscles and bones of the trunk [87].

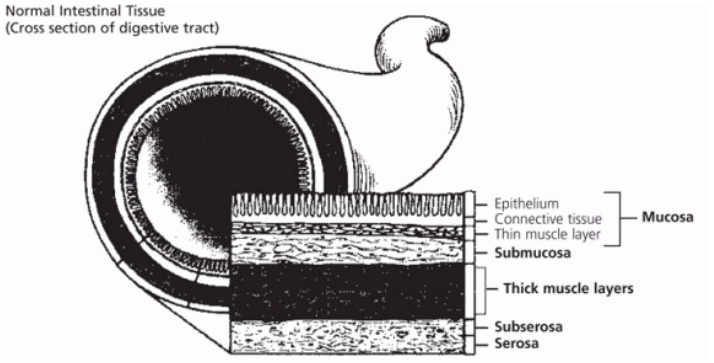



Figure 3.: The layers of the bowel wall [1].

## 1.2 THE COLORECTAL CANCER

### 1.2.1 *Incidence*

The most common cancer of the digestive system is CRC, also known as bowel cancer. It is a term for cancer that starts in the colon or rectum. About 70 percent of the CRC cases occur in the colon and about 30 percent in the rectum [95]. According to the latest worldwide cancer statistics (2012), the CRC is the third most frequent cancer worldwide (Figure 4a) and the fourth deadliest (Figure 4b), for both sexes. Almost 55% of the cases

occur in more developed regions. As shown in Figure 5, the highest estimated rates are in Australia and New Zealand and the lowest are in Western Africa. CRC mortality is low when compared to other cancers (694,000 deaths, 8.5% of the total) and the majority of CRC deaths occurs (52%) in the less developed regions of the world. It is estimated that, in 2020, the world will have 1.7 million new CRC cases and almost 855 thousand CRC deaths [29].

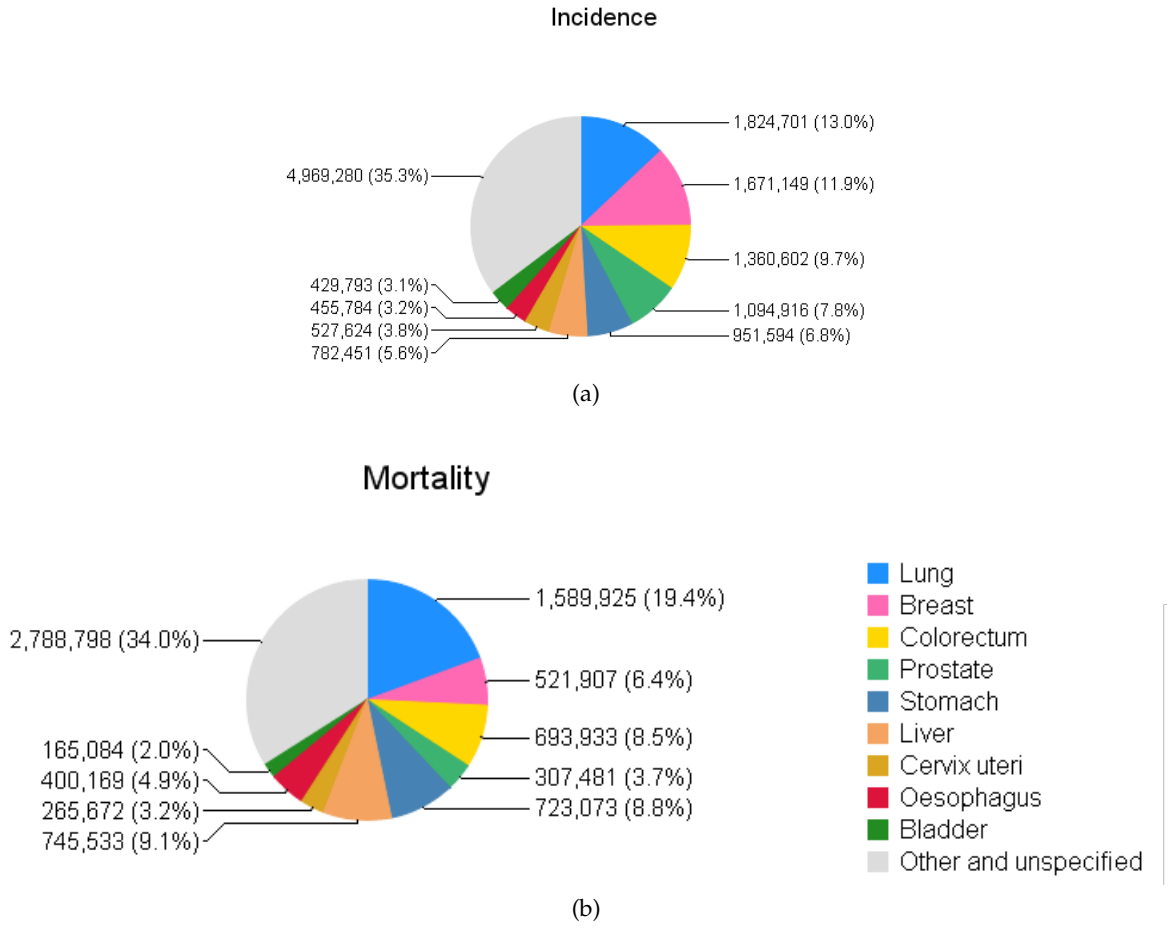



Figure 4.: Estimated cancer incidence and mortality worldwide in 2012, for both sexes [29].

Most CRCs begin as a small growth called a polyp (Figure 6). This growth is a benign tumor and not all can become into cancer. It starts in the inner lining of the colon or rectum and grows toward the center, and this process can take many years. Taking out a polyp early, when it is small, may keep it from becoming cancer [95].

More than 95% of all large bowel tumors are adenocarcinomas (tumors which start in the gland cells in the lining of the bowel wall). The gland cells produce mucus to lubricate the inside of the colon and rectum (that makes it easier for the stool to pass through the bowel). Other rare types include lymphoma and squamous cell carcinoma [1].

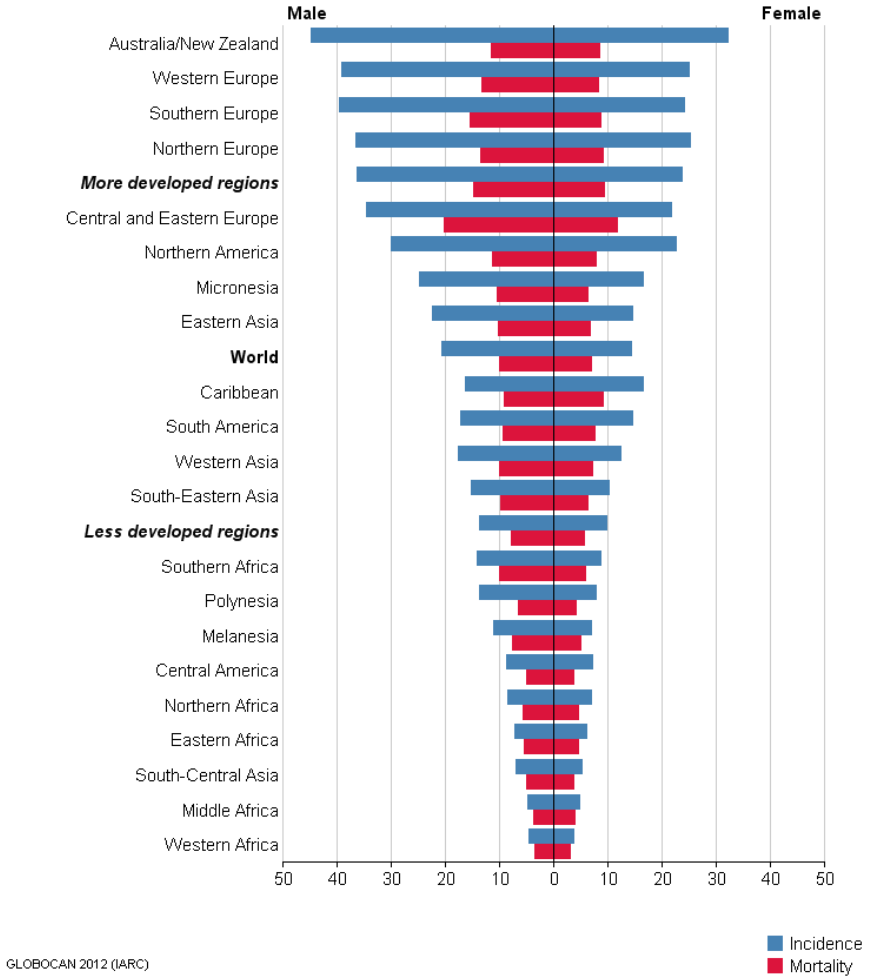



Figure 5.: Estimated cancer age-standardised rates worldwide in 2012, per 100.000 people [29].

The exact cause of CRC is unknown. However, at least eight different genes can be traced to dietary fat, particularly animal fat [95].

The risk of CRC increases with age. Individuals with a personal or family history of CRC or polyps, inherited CRC syndromes (i.e., familial adenomatous polyposis and hereditary nonpolyposis CRC), and patients with ulcerative colitis or Crohn's disease are at higher risk, and thus may require screening at an earlier age than the general population. Modifiable factors associated with increased risk include a diet high in fat and red or processed meat, but low in fiber, low calcium intake, high caloric intake, physical inactivity, and obesity. In addition, smoking and excessive alcohol intake may play a role in CRC development [95, 4, 57].

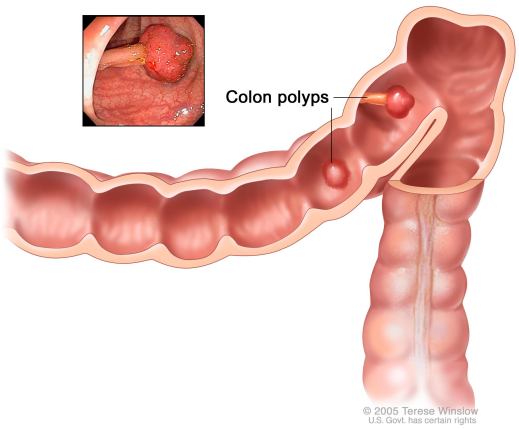



Figure 6.: Polyps in the colon: some polyps have a stalk and others do not (the photo in the figure shows a polyp with a stalk) [57].

### 1.2.2 *Symptomatology and Screening*

Symptoms may include blood seen in the stool (can be bright red or very dark), unexplained persistent constipation alternating with diarrhea, changes in the diameter of stool, intermittent abdominal pain and the feeling of inadequate emptying of the bowel [95, 4, 57].

CRC can largely be prevented by the detection and removal of adenomatous polyps (precancerous polyps). From 2006 to 2007, the American Cancer Society, the US Multi Society Task Force on Colorectal Cancer (a consortium representing the American College of Gastroenterology, the American Society of Gastrointestinal Endoscopy, the American Gastroenterological Association, and the American College of Physicians), and the American College of Radiology came together to develop consensus guidelines for CRC screening. In a range of options for CRC screening, the guidelines distinguish between two general categories, according to current technology: stool tests – which include tests for occult blood or exfoliated DNA – and structural exams – which include flexible sigmoidoscopy (FSIG), colonoscopy, double-contrast barium enema (DCBE), and computed tomographic colonography (CTC) [51]. The screening tests recommended for CRC screening in men and women aged 50 or older at average risk are summarized in Table 1 (the complexity involves patient preparation, inconvenience, facilities and equipment needed, and patient discomfort) [90].

### 1.2.3 *Staging*

Cancer staging is the process of finding out how widespread a cancer is, determining how much cancer is in the body and where it is located. Staging describes the severity of a individual's cancer based on the size and/or extent of the original (primary) tumor

Table 1.: Recommendations for CRC screening [90].

| Test | Benefits | Performance & Complexity | Limitations | Test Time Interval |
|---|---|---|---|---|
| Flexible Sigmoidoscopy | Fairly quick;<br>Few complications;<br>Minimal bowel preparation;<br>Minimal discomfort;<br>Does not require sedation or a specialist. | Performance: High for rectum & lower one-third of the colon;<br>Complexity: Intermediate | Views only one-third of colon;<br>Bowel preparation needed;<br>Cannot remove large polyps;<br>Small risk of infection or bowel tear;<br>Slightly more effective when combined with annual fecal occult blood testing;<br>Colonoscopy necessary if abnormalities are detected. | 5 years |
| Colonoscopy | Examines entire colon;<br>Can biopsy and remove polyps;<br>Can diagnose other diseases;<br>Required for abnormal results from all other tests. | Performance: Highest;<br>Complexity: Highest. | Can miss some polyps and cancers;<br>Full bowel preparation needed;<br>Can be expensive;<br>Sedation of some kind usually needed, necessitating a chaperone;<br>Patient may miss a day of work;<br>Highest risk of bowel tears or infections compared to other tests. | 10 years |
| Double-contrast Barium Enema | Can usually view entire colon;<br>Few complications;<br>No sedation needed. | Performance: High;<br>Complexity: High. | Can miss some small polyps and cancers;<br>Full bowel preparation needed;<br>Cannot remove polyps;<br>Exposure to low-dose radiation;<br>Colonoscopy necessary if abnormalities are detected. | 5 years |
| Computed Tomographic Colonography | Examines entire colon;<br>Fairly quick;<br>Few complications;<br>No sedation needed;<br>Noninvasive. | Performance: High;<br>Complexity: Intermediate. | Can miss some polyps and cancers;<br>Full bowel preparation needed;<br>Cannot remove polyps;<br>Exposure to low-dose radiation;<br>Colonoscopy necessary if abnormalities are detected. | 5 years |
| Fecal Occult Blood Test | No bowel preparation;<br>Sampling is done at home;<br>Low cost;<br>Noninvasive. | Performance: Intermediate for cancer;<br>Complexity: Lowest. | May require multiple stool samples;<br>Will miss most polyps and some cancers;<br>Higher rate of false-positives than other tests;<br>Pre-test dietary limitations;<br>Slightly more effective when combined with a flexible sigmoidoscopy every five years;<br>Colonoscopy necessary if abnormalities are detected. | Annual |
| Stool DNA Test | No bowel preparation;<br>Sampling is done at home;<br>Requires only a single stool sample;<br>Noninvasive. | Performance: Intermediate for cancer;<br>Complexity: Low. | Will miss most polyps and some cancers;<br>High cost compared to other stool tests;<br>New technology with uncertain interval between testing;<br>Colonoscopy necessary if abnormalities are detected. | Uncertain |

and whether or not cancer has spread in the body [95]. It is performed for diagnostic and research purposes, and to help the doctor plan the appropriate treatment. It also gives a common terminology for evaluating the results of clinical trials and comparing the results of different trials. On the one hand, if the stage is based on the results of the physical exam, biopsy, and any imaging tests, it is called a clinical stage. On the other hand, when it is performed a surgery or biopsy, the results can be combined with the factors used for the clinical stage, determining the pathologic stage. A cancer is always referred to by the stage it was given at diagnosis, even if it gets worse or spreads. The survival statistics and information on treatment by stage for specific cancer types are based on the original cancer stage at diagnosis.

One of the most widely used cancer staging systems is the TNM (for tumors/nodes/metastases) system, from the *American Joint Committee on Cancer (AJCC)*. It is based on the size and/or extent of the primary tumor (T) – see Figure 7 –, the amount of spread to nearby lymph nodes (N), and the presence of metastasis (M) or secondary tumors formed by the spread of cancer cells to other parts of the body. The TNM system assigns a number to

each letter to indicate the size and/or extent of the primary tumor and the degree of cancer spread. It is worth noting that each cancer type has its own classification system: letters and numbers do not always mean the same thing for every kind of cancer. Table 2 shows all the definitions for T, N, and M.

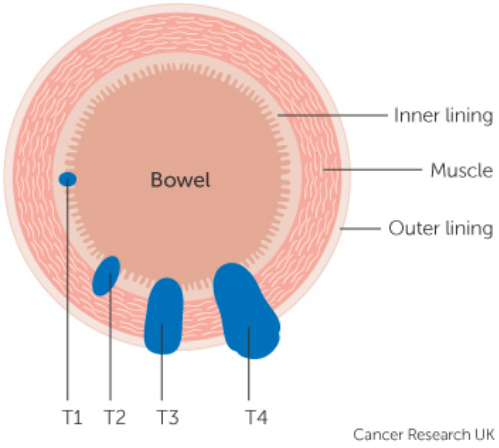



Figure 7.: The degree of invasion of the intestinal wall [15].

By combining the TNM information, it is possible to obtain an overall "Stage". It is expressed in roman numerals: from stage 0 (the least advanced) to stage IV (the most advanced), as shown in Figure8. These stages can be subdivided using letters, as for instance IIIA and IIIB.

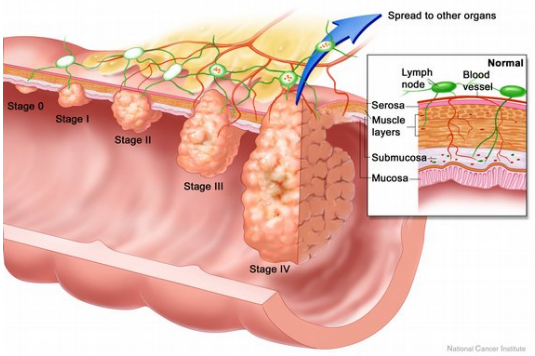



Figure 8.: Grouping of TNM classification for colorectal cancer.

In addition to the TNM staging system, there are other common staging schemes in use. The Dukes' staging classification was originally published by Cuthbert E. Dukes in 1932 for rectal cancer only and does not include distant metastases. In 1949 it was adapted by Kirklin and later (in 1953) by Astler and Coller for colon and rectum. To include stage for unresectable tumors and distant metastases, it was revised by Turnbull in 1967. Astler-Coller

Table 2.: Classification of colorectal cancers according to local invasion depth (T stage), lymph node involvement (N stage), and presence of distant metastases (M stage) [26].

| | |
|---|---|
| **Primary Tumor (T)** | |
| TX | Primary tumor cannot be assessed |
| To | No evidence of primary tumor |
| Tis | Carcinoma *in situ*: intraepithelial or invasion of lamina propria |
| T1 | Tumor invades submucosa |
| T2 | Tumor invades muscularis propria |
| T3 | Tumor invades through the muscularis propria into the pericolorectal tissues |
| T4a | Tumor penetrates to the surface of the visceral peritoneum |
| T4b | Tumor directly invades or is adherent to other organs or structures |
| **Regional Lymph Nodes (N)** | |
| NX | Regional lymph nodes cannot be assessed |
| No | No regional lymph node metastasis |
| N1 | Metastasis in 1-3 regional lymph nodes |
| N1a | Metastasis in one regional lymph node |
| N1b | Metastasis in 2-3 regional lymph nodes |
| N1c | Tumor deposit(s) in the subserosa, mesentery, or nonperitonealized pericolic or perirectal tissues without regional nodal metastasis |
| N2 | Metastasis in four or more regional lymph nodes |
| N2a | Metastasis in 4-6 regional lymph nodes |
| N2b | Metastasis in seven or more regional lymph nodes |
| **Distant Metastasis (M)** | |
| Mo | No distant metastasis |
| M1 | Distant metastasis |
| M1a | Metastasis confined to one organ or site (eg, liver, lung, ovary, nonregional node) |
| M1b | Metastases in more than one organ/site or the peritoneum |

and Turnbull stagings are also sometimes called Dukes or modified Astler-Coller (MAC) [95]. It is possible to observe in Table 3 that these staging systems and the correspondence between them.

Treatment for CRC depends on several factors, including the type and stage of cancer. Early stages of CRC are often treated with surgery – 95% of Stage I and 65-80% of Stage II –, extracting the cancer from the colon, rectum or even in other organs in the body where the cancer has spread to. Radiation therapy, i.e., applying high-energy x-rays to destroy cancer cells, may be required to minimize the recurrence risk in rectal cancer. Other types of treatment that are often used are chemotherapy or targeted therapy. Chemotherapy is a type of cancer treatment that uses medication (chemicals) to neutralize cancer cells, usually by stopping the ability that cancer cells have to grow and divide. These chemicals can be

Table 3.: Anatomic Stage/Prognostic Groups, seventh edition [26, 42, 35].

| Stage | T | N | M | Dukes | MAC |
|---|---|---|---|---|---|
| 0 | Tis | N0 | M0 | - | - |
| I | T1 | N0 | M0 | A | A |
| | T2 | N0 | M0 | A | B1 |
| IIA | T3 | N0 | M0 | B | B2 |
| IIB | T4a | N0 | M0 | B | B2 |
| IIC | T4b | N0 | M0 | B | B3 |
| IIIA | T1-T2 | N1/N1c | M0 | C | C1 |
| | T1 | N2a | M0 | C | C1 |
| IIIB | T3-T4a | N1/N1c | M0 | C | C2 |
| | T2-T3 | N2a | M0 | C | C1/C2 |
| | T1-T2 | N2b | M0 | C | C1 |
| IIIC | T4a | N2a | M0 | C | C2 |
| | T3-T4a | N2b | M0 | C | C2 |
| | T4b | N1-N2 | M0 | C | C3 |
| IVA | Any T | Any N | M1a | D | D |
| IVB | Any T | Any N | M1b | D | D |

injected into a vein or given by mouth, injected directly into an artery leading to a part of the body containing a tumor or can even given directly into the hepatic artery. Targeted therapy is also a treatment that uses drugs. However, it is different from traditional chemotherapy. This treatment has as targets the specific genes of the cancer, proteins, or surrounding tissues that contribute to the cancer growth and survival. Targeted therapy typically has less severe side effects. It could be used either along with chemotherapy or by itself, if chemotherapy is no longer working. Depending on the stage of the cancer, two or more types of therapy may be combined at the same time or used sequentially. When cancer has spread away from the original tumour site (stage IV), most often it cannot be cured. However, the cancer may be treatable and its growth and symptoms could be managed [90].

Based on largest population-based cancer registry in the United States, the *Surveillance, Epidemiology, and End Results (SEER)* database – provided by the National Cancer Institute –, from 2005 to 2011, the five- and ten-year relative survival rates, i.e., statistics that compare the survival of patients diagnosed with CRC with the survival of people in the general population (with the same age, race, and sex and who have not been diagnosed with cancer), are 65% and 58%, respectively. Survival rates for CRC depend on multiple factors, they often include the stage (Figuer 9). When CRC is detected at a localized stage (cancer is only in the part of the body where it started), the five-year survival is 90.1%. If the cancer has spread to a different part of the body, to nearby organs or lymph nodes, the five-year survival drops to 71%. If the disease has spread to distant organs, the five-year survival

rate is 13%. The five-year survival rate for patients who have just one or a few tumors that have spread (for instance, to the lung or liver), can be improved if surgical removal of these tumors is able to eliminate the cancer [4].

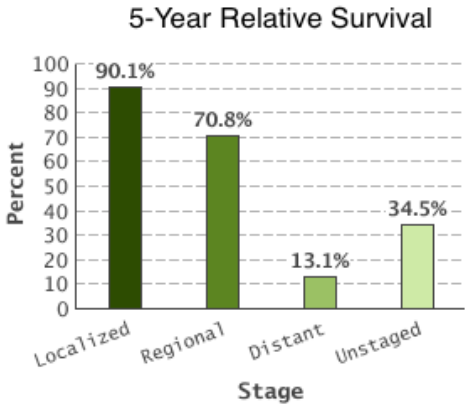



Figure 9.: 5-Year Relative Survival (from SEER 18 2005-2011, All Races, Both Sexes by SEER Summary Stage 2000).

## 1.3 MOTIVATION

The term *Artificial Intelligence (AI)* was coined be Jon McCarthy in 1956. He defined it as "the science and engineering of making intelligent machines" at a conference at the campus of Dartmouth College. This science, when used in medical applications, is called *Artificial Intelligence in Medicine (AIM)*. The earliest work in AIM dates to 15 years after AI was founded [66, 93]. Called "The DENDRAL Project", it was a cooperative work, which brought together computer scientists, chemists, geneticists and philosophers of science, to show the capacity to represent and apply expert knowledge in symbolic form [53].

The ability to learn is viewed as the typical characteristic of an intelligent being. Consequently, to develop devices that can get skills from experiences has been one of the driving ambitions of AI. *Machine Learning (ML)* is another core part of AI. Its developments have resulted in a set of techniques which have the potential to alter the way in which knowledge is created [20, 62].

Data mining is defined as the automatic search, in large amounts of data, for patterns. It is also known as knowledge-discovery in databases and uses computational techniques from statistics, ML and pattern recognition.

Medicine is the practice of the diagnosis, treatment and prevention of disease, and the promotion of health. It is a critical area where the time can be crucial. For shortage of time, most medical decisions are based on quick judgments and depend on the memory

of the physician [93]. Training and recertification procedures may improve the physician skills, encouraging him to keep more of the relevant information in mind [23, 24]. However, fundamental limitations of human memory and recall mechanisms, coupled with the exponential growth in knowledge, mean that most of what is known cannot be seized by most individuals. To overcome this situation, by helping to organize, store, and retrieve appropriate medical knowledge needed, some new areas emerged, namely eHealth, Clinical Decision Support Systems, Computer-Interpretable Guidelines and Reasoning under Uncertainty [63, 61, 52]. Some of these technologies are able to suggest appropriate diagnostic, prognostic and therapeutic decisions.

Accurate prediction of survival is one of the most interesting and challenging tasks for physicians and it is important for various purposes, such as medical decision making, patient counselling and benchmarking [49, 99, 16]. The level of physician experience to estimate the survival might affect how prognostic is formulated. However, even an experienced oncologist has difficulty to predict accurately survival time of a patient with cancer [94].

Survival statistics could help estimate the prognosis of patients, but they are based on large groups of people, they cannot be used to predict exactly what will happen to an individual patient [58]. Kaplan-Meier is one of most frequently used method in the conventional analysis of survival problems [99]. It is the simplest way of computing the survival over time, can be calculated for two groups of subjects and it involves the computing of probabilities of occurrence of event (death) at a certain point of time [34].

In order to exploit the implied knowledge in large clinical datasets, some sophisticated modeling in AI approaches to medical reasoning have been exploited through ML techniques [93]. These techniques have competence to discover and identify patterns and the relationships between them, from complex datasets [49]. Based on this, herein the development of a survival prediction model is proposed for CRC patients. To develop the prediction model, ML will be used to discover the relationships between the different variables and their weights in survival prediction. Approaching of the mobile health (mHealth), the model will be available in different platforms (smartphones and tablets) and it will target health care professionals. The developed tool will employ current technologies related with web development, ubiquitous computing and intelligent interfaces.

## 1.4 OBJECTIVES

This dissertation project has the main goal of developing a model to predict the survival of patients with colon and rectal cancer. Several points were delineated to achieve an appropriate solution able to help physicians improve their practice, such as:

1. Employ machine learning techniques to process the collected information of CRC patients from a database;

2. Construct an accurate model able to predict the survival at 1-, 2-, 3-, 4- and 5-years after the diagnosis and treatment;
3. Construct an accurate model able to predict the conditional survival at patients who had already survived at 1-, 2-, 3- and 4-years after the diagnosis and treatment;
4. Find the most relevant features to construct the models through a feature selection process by a software;
5. Compare the features selected by a software with the features that physicians consider most relevant;
6. Ascertain whether characteristics used to predict the survival for colon and rectal cancer patients are the same;
7. Determine if the existing models are effective and are available to health care professionals to evaluate the developed models;
8. Develop a cross platform mobile to make available the models to health care professionals;
9. Construct an online learning service to recalculate the models after several entries in the application.

## 1.5 DOCUMENT STRUCTURE

The present work is constituted by six chapters, structured as follows:

**INTRODUCTION** In the first chapter are introduced important concepts to the comprehension of all the work. Is made a description of the disease and the work is framed. The motivation, objectives and document structure are presented.

**STATE OF THE ART** The second chapter contains the actual solutions for prediction of survivability in colorectal cancer patients. It presents the survivability prediction tools, for colon and rectal cancer, and prediction models which are not available in any form to users. A discussion of the state of the art is made at the end.

**DEVELOPMENT OF THE PREDICTION MODEL** The third chapter, is the main chapter of the dissertation. It describes all the processes of development of the prediction model, from the raw data to modeling and evaluation, including the testing phase.

**EXPERIMENTAL RESULTS** The fourth chapter reveals the results from the development of the prediction model. All the developed survival and conditional survival models, for

colon and rectal cancer patients, are compared using metrics and the best models are selected to embed a prediction tool. A discussion is made at the end of the chapter.

**DEVELOPMENT OF AN APPLICATION** The fifth chapter describes the processes of development of an application, in order to make available the models to physicians. The gathering of functionals requirements and non-functionals requirements is made. Also, the architecture and the interface of the tool is shown and a use case is made.

**CONCLUSIONS AND FUTURE WORK** Finally, the sixth and last chapter of this dissertation synthesizes all the accomplished work and the main conclusions from it. A prospect for future work is mentioned.

# 2

# STATE OF THE ART

In this chapter the most relevant related work within the context of this thesis is presented. It is separated in two main sections: one for *Colon Cancer (CC)* and another for *Rectal Cancer (RC)*. In each of the main sections a review of the current tools to predict the survival in patients with CC or RC is made. In the section regarding colon cancer, are also reported some survival models developed, for these cancers, which are not available in any form to users. Ultimately, in a third section, is made a discussion of the state of the art.

## 2.1 SURVIVABILITY PREDICTION TOOLS

### 2.1.1 *For Colon Cancer*

*Bush and Michaelson (2009)*

The *CancerMath.net* group – a section of the Laboratory for Quantitative Medicine from Massachusetts – developed a series of web-based calculators[1] for accurately predicting the clinical outcome for individual cancer patients. These tools are available for melanoma, breast, renal, colon, head and neck cancers.

The CC outcome calculator is a work in progress and it was reported by Bush and Michaelson [14]. This tool provides information on survival expectation at the time of diagnosis, for each of the first 15 years after diagnosis. Also provided is the life expectancy with and without cancer, and the reduction of life expectancy caused by cancer.

The data used to developed the CC tool was extracted from the SEER dataset, from 1973 to 2006.

The variables used to construct the model were: the age of the patient at diagnosis, gender (male or female), tumor diameter (in cm), number of positive nodes, carcinoembryonic antigen (CEA)[2] status (positive or negative), histological type, grade (well differentiated, moderately differentiated, poorly differentiated or undifferentiated), site (region of

1 This tool is available at http://www.lifemath.net/cancer/coloncancer/outcome/index.php.

2 CEA is a glycoprotein and is used as a tumor marker. In increased large are associated with adenocarcinoma, especially colorectal cancer.

the colon) and farthest tumor extension. Through the information of farthest tumor extension and number of positive nodes, this tool can provide the TNM classification and AJCC stage group.

*Nodes + Prognostic Factors (NAP)* method was created to more accurately model the CC lethality based on the number of positive nodes, combined with prognostic factors. The NAP method is written below, in equations (1), (2) and (3). The prognostic factors are represented by $(g_1, g_2, g_3, \ldots, g_n)$. Q, R and $j_{primary}$ are empirically derived constants.

$$L_{primary} = 1 - e^{(-Q*(g_1*g_2*g_3*\ldots*gn)*j_{primary})} \tag{1}$$

$$L_{nodes} = 1 - e^{(-R*(\#\,positive\,nodes))} \tag{2}$$

$$L_{overall} = L_{primary} + L_{nodes} - (L_{primary} * L_{nodes}) \tag{3}$$

For the CC model, the $j_{primary}$ constant of 0.61299589 was only applied to tumors with zero known positive nodes. For tumors with any positive nodes, or unknown positive nodes, the $j_{primary}$ is 1.

Figure 10 shows that the 3-, 5- and 15-year Kaplan-Meier disease-specific death rates. It shows an approximately 25% reduction in deaths, from the 1970s to 2003.

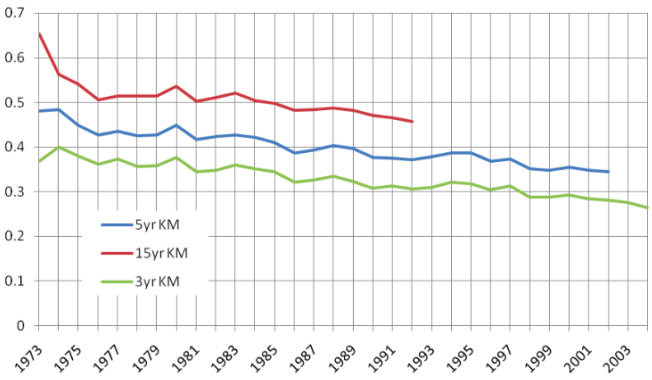



Figure 10.: Disease-specific Kaplan-Meier lethality by year (from 1973 to 2003) [14].

The interface of this tool is shown in Figure 11. No performance results are known for this tool.

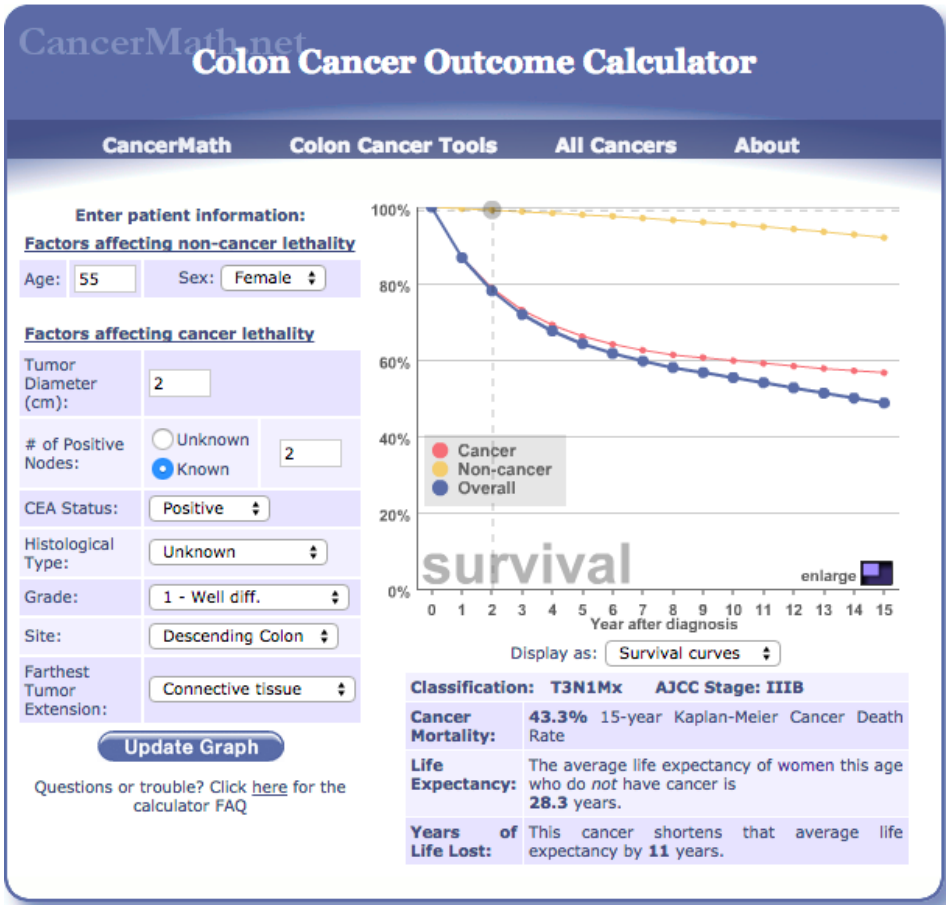



Figure 11.: Web-based calculators for accurately predicting the clinical outcome for individual colon cancer patients and results [13].

*Chang et al. (2009)*

From The University of Texas M. D. Anderson Cancer Center, Chang et al. [18] created a browser-based calculator[3] to predict individualized disease-specific survival and conditional survival[4].

By utilizing data from the SEER registry, 83,419 patients with colon adenocarcinoma diagnosed between 1988 and 2000 were analyzed.

The variables used to develop the model were: age of the patient at diagnosis (categorized into age less than 50 years, 50 to 59 years, 60 to 69 years, 70 to 79 years and $\geq$ 80 years), gender (male or female), ethnicity (white, black or other), tumor grade (well differentiated, moderately differentiated, and poorly differentiated or undifferentiated) and AJCC sixth edition stage group. The inclusion of the SEER region, year of diagnosis, marital status

3 The browser-based conditional survival calculator is available at http://www3.mdanderson.org/coloncalculator.
4 The survival probability calculated after a given length of survival, including only individuals who have survived to a predefined time of interest.

and tumor location was also tested, but these features did not improve model performance or prediction accuracy. A multivariate Cox regression analysis was performed by using the Breslow method for ties to evaluate the simultaneous effect of multiple variables on survival.

The Kaplan-Meier unadjusted (10-year) disease-specific survival probabilities for patients diagnosed with CC stratified by AJCC stage (sixth edition) is shown in Figure 12.

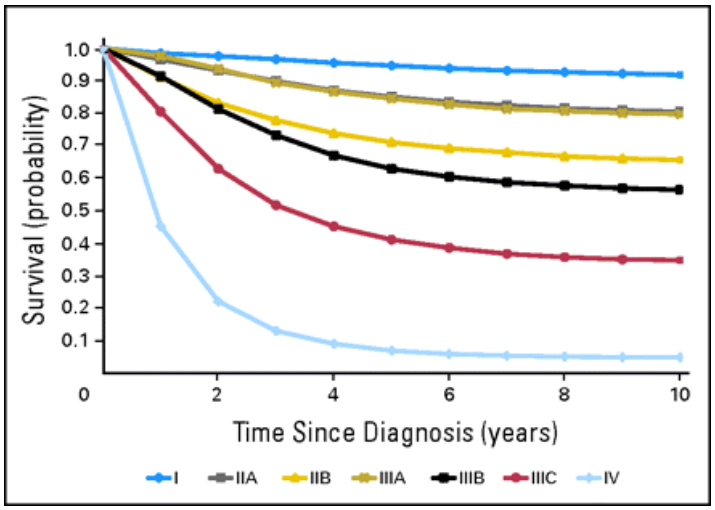



Figure 12.: Unadjusted AJCC disease-specific survival for patients with CC who were diagnosed between 1988 and 2000 [18].

The conditional survival was applied to obtain a more accurate survival probability. It is utilized especially when the initial prognosis is poor. Conditional survival estimates were calculated by using the multiplicative law of probability after adjustment for variables. The *Concordance Index (C-index)* of this implementation was 0.816.

Figure 13 shows the interface of this tool and Figure 14 shows the results.

*Weiser et al. (2011)*

From Memorial Sloan Kettering Cancer Center, Weiser et al. [104] developed a tool[5] with the ability to predict the overall survival probability of the CC patient at least five years following surgical removal of all cancerous tissue.

The tool can produce three different estimates based on the amount of data included, and the accuracy increases with the amount of submitting information. For that purpose, three nomograms using multivariable regression with Cox proportional hazards modeling were created. This tool also provides a highly likely range for the probability of survival, known as the 95% confidence interval.

5 This tool is available at https://www.mskcc.org/nomograms/colorectal/overall-survival-probability.

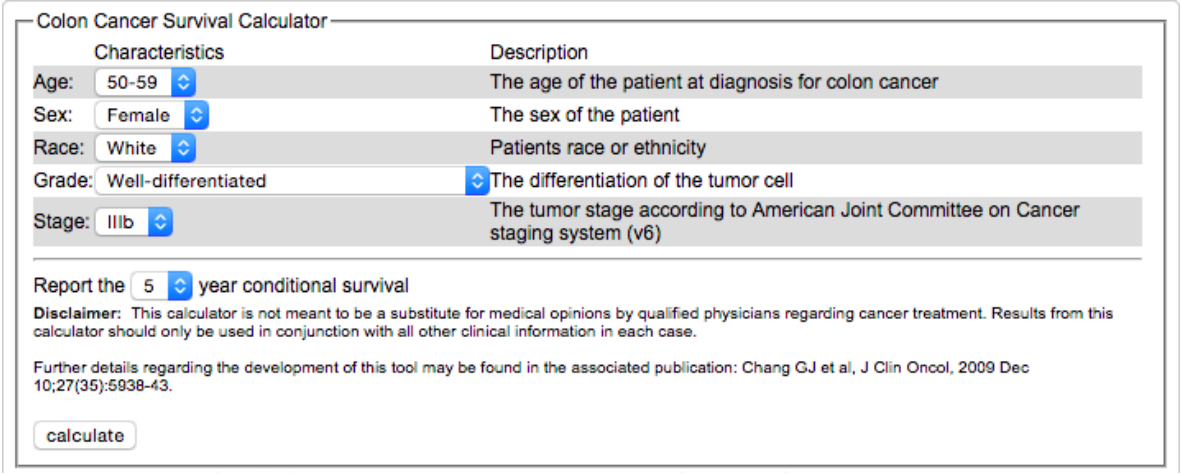



Figure 13.: Browser-based calculator to predict individualized disease-specific survival and conditional survival for colon cancer patients [17].

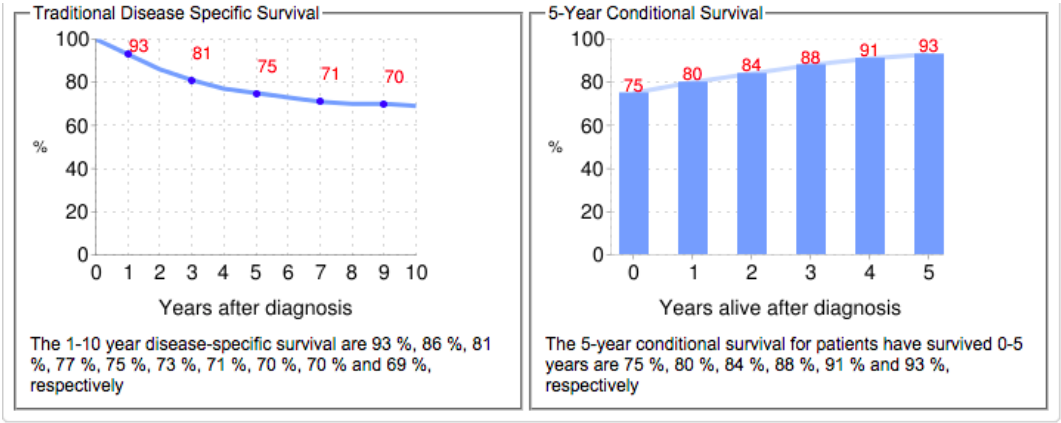



Figure 14.: Results of the browser-based calculator to predict individualized disease-specific survival and conditional survival for colon cancer patients [17].

To construct and validate the three survival models the records from 128,853 primary colon cancer patients reported to the SEER from 1994 to 2005 were applied.

For a basic estimate of overall survival probability it is necessary to know the depth of tumor penetration into the colon wall (T stage) and the N stage, according to the TNM anatomic staging system, introduced in Section 1.1. For a more accurate estimate, it is necessary to know T stage, the number of positive lymph nodes (value between 0 and 16) and the number of total lymph nodes (value between 0 and 45). For the most accurate estimate, in addition to the data required in the previous accurate estimate, it vital know the age of the patient at the time of surgery, gender (male or female) and tumor differentiation/grade (poor, moderate or well differentiated).

All these variables were chosen *a priori* on the basis of their well-established independent associations with overall survival and their availability in the SEER registry.

Kaplan-Meier overall survival curves for the entire population, according to the AJCC classification schema (seventh edition), are shown in Figure 15.

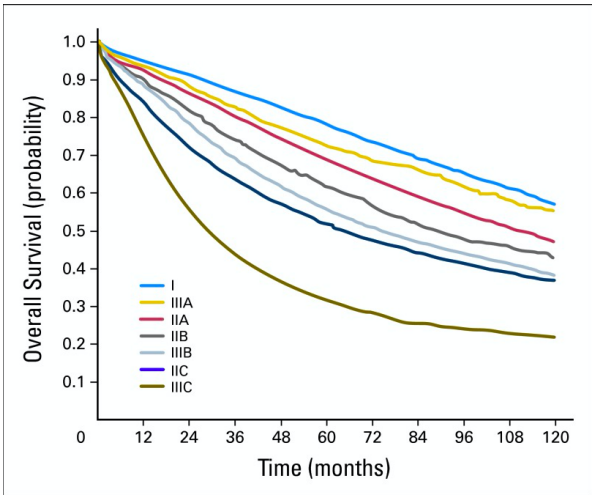



Figure 15.: Kaplan-Meier overall survival on the basis of the seventh edition of the AJCC Staging Manual [104].

The simplest nomogram, based only on T and N elements, presented the minor C-index, with 0.61 (95% CI, 0.60 to 0.62). It was followed by the model that includes the number of lymph nodes examined and number of metastatic lymph nodes examined, with 0.63 (95% CI, 0.62 to 0.64). Finally, the highest C-index belongs to the model including the pathologic tumor differentiation and demographic variables of age and gender, with 0.68 (95% CI, 0.67 to 0.68). The *Receiver Operating Characteristic (ROC)* curves for the extended model had higher sensitivity, at all values of specificity, than the TNM system and calibration curves indicated no deviation from the reference line.

The interface of this tool is shown in Figure 16 and the results in Figure 17.

*Renfro et al. (2014)*

From Mayo Clinic, Renfro et al. [86] created a clinical calculator[6] for overall survival and time to recurrence for stage III colon cancer. It was developed in order to obtain predicted probabilities of being recurrence-free at three years and alive at five years over the start of treatment, with confidence intervals.

Multivariable Cox proportional hazards models for overall survival and time to recurrence were formulated using data from 15,936 stage III CC patients. These data were col-

6 This clinical calculator is available for use at http://www.mayoclinic.org/medical-professionals/cancer-prediction-tools/colon-cancer.

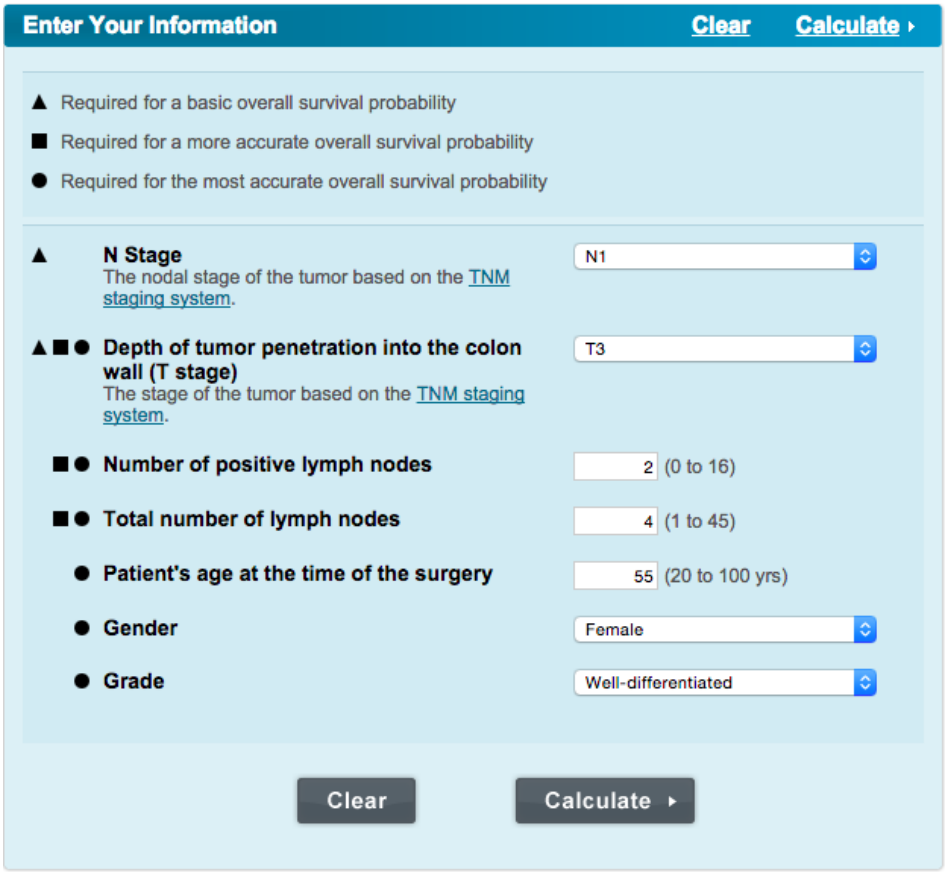



Figure 16.: Colorectal cancer nomogram: overall survival probability [103].

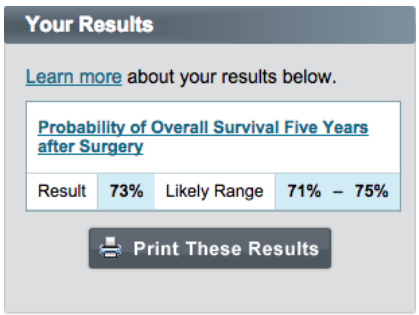
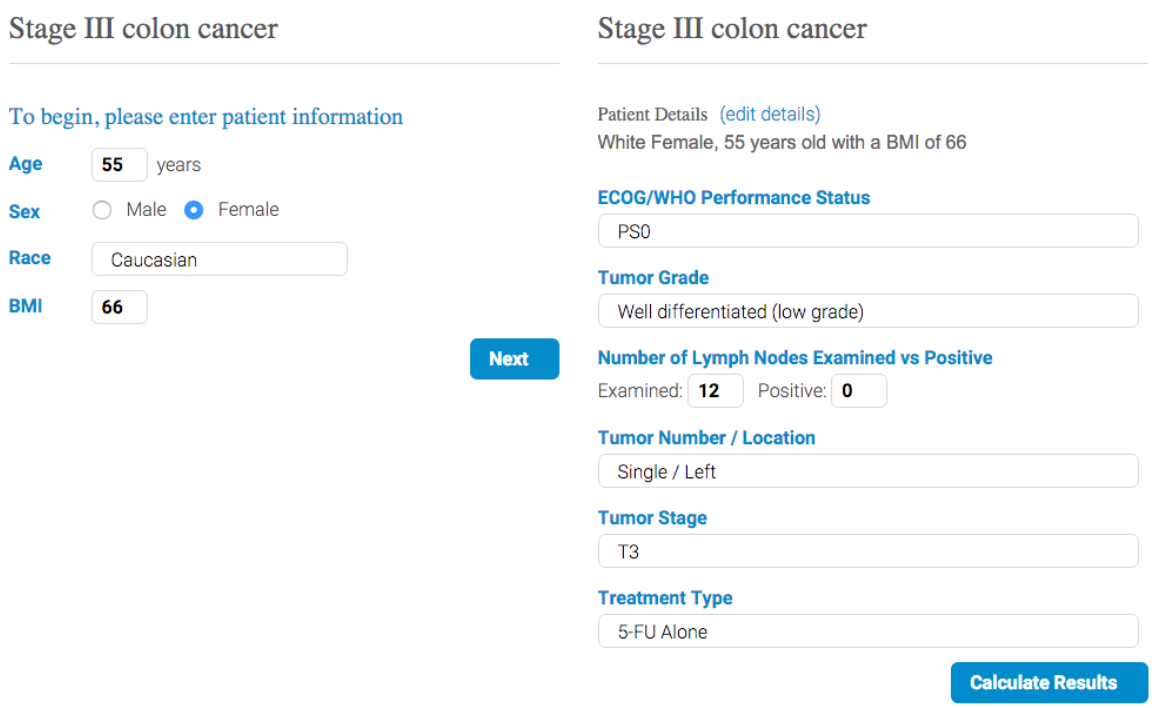



Figure 17.: Results of colorectal cancer nomogram [103].

lected from 12 randomized clinical trials, from 1989 to 2002, contained in the *Adjuvant Colon Cancer End Points (ACCENT)* database.

Models were constructed using variables such as the age (as continuous variable), sex (male or female), race (white, black, asian, or other), body mass index (as a continuous variable), Eastern Cooperative Oncology Group/World Health Organisation performance




(a) Patient characteristics: first screen.

(b) Patient characteristics: second screen.

Figure 18.: Web calculator to predict recurrence and overall survival in stage III colon cancer [85].

status scale[7] [60] (0, 1, 2+), tumor grade (1, 2, 3+), tumor stage (T-stage; T1, T2, T3, T4), ratio of the number of positive lymph nodes to the number of nodes examined (as continuous variable, between 0 and 1), number and location of primary tumors (any multiple, single left, single right, or single transverse/flexures), and treatment class (oral/infusional/bolus 5FU variations vs 5FU with oxaliplatin vs 5FU with irinotecan).

Model for overall survival had a C-index of 0.66. Figure 18 shows the interface of this tool and Figure 19 its results.

### 2.1.2 *For Rectal Cancer*

*Wang et al. (2011)*

From the OHSU Knight Cancer Institute, Wang et al. [102] developed an interactive tool[8] to make an individualized prediction of the conditional survival for a RC patient, after a certain period of time passed since diagnosis and treatment.

7 Scale developed by the Eastern Cooperative Oncology Group (ECOG), part of the ECOG-ACRIN Cancer Research Group. It describes a patient's level of functioning in terms of their ability to care for themself, daily activity, and physical ability (walking, working, etc.).

8 This prediction calculator is available at http://skynet.ohsu.edu/nomograms/gastrointestinal/rectal.php.

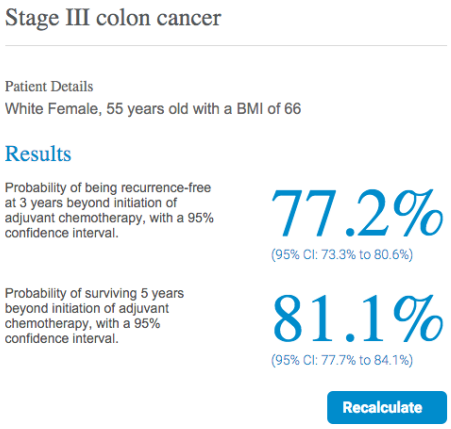



Figure 19.: Results of the web calculator to predict recurrence and overall survival in stage III colon cancer [85].

The prediction calculator was constructed based on data from 42,830 RC patients who were diagnosed between 1994-2003, from the SEER 17 database. Conditional survival prediction is calculated from a Cox proportional hazards model.

The primary outcome variable was overall survival conditional on having survived up to 5 years from diagnosis. Covariates included in the model were age (as a continuous variable), race (white, black, Asian/Pacific Islander, Alaskan/American Indian), sex, and stage (AJCC TNM grouped stage from third edition).

The 10-year actuarial survival data (Figure 20) were used to calculate the 5-year observed conditional survival in categories of stage, age, gender, and race.

The C-indexfor the model of this approach was 0.75. Figure 21 shows a screenshot of this interactive web-based prediction tool.

*Valentini et al. (2011)*

From the MAASTRO Clinic, Valentini et al. [98] developed nomograms[9] as a tool to predict the probability that a rectal cancer patient will be alive or will have local recurrence or distant metastasis after delivery of long-course radiotherapy, with optional concomitant and/or adjuvant chemotherapy, over a 5-year period after surgery.

Based on Cox regression, multivariate nomograms were developed. They were built based on 2,795 individual patient data collected from five European randomized trials[10] that tested preoperative chemoradiotherapy against preoperative radiotherapy or postoperative chemoradiotherapy and adjuvant chemotherapy, between 1992 to 2003.

9 This prediction calculator is available at http://www.predictcancer.org/.

10 Trial name: European Organisation for Research and Treatment of Cancer, *Fédération Francophone de Cancérologie Digestive*, Working Group of Surgical Oncology/Working Group of Radiation Oncology/Working Group of Medical Oncology of the Germany Cancer Society, Polish and Italian.

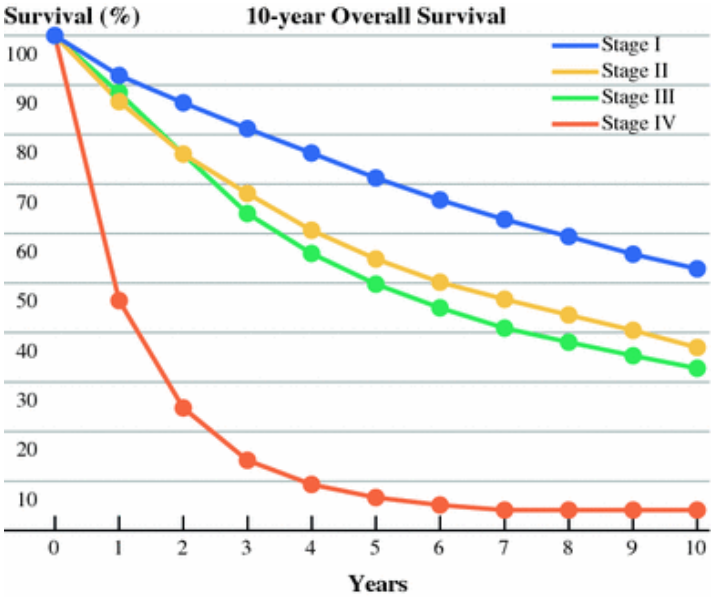



Figure 20.: 10-year Kaplan–Meier overall survival by stage [102].

Training the Cox model (training data), important predictors for outcome and the variables that have a significant effect, were selected. The required information for the overall survival calculator was gender (male or female), age at the date of randomization (as a continuous variable), clinical tumor stage (1+2, 3 or 4), radiotherapy dose ($<$ 45 Gy, 45 Gy, and $>$ 45 Gy), surgery procedure (low anterior resection [LAR] or abdominoperineal resection [APR]), adjuvant chemotherapy (yes/no), pathological tumor and nodal stage. The concomittant chemotherapy (yes/no) is used to calculate the local recurrence. However, it must be inserted, even for overall survival prediction, because it is a field required for the tool.

Kaplan-Meier curves of risk group stratification for overall survival for dataset validation can be observed in Figure 22.

The nomogram for overall survival had a C-index of 0.70 (95% CI, 0.65 to 0.74). Figure 23 shows the interface of this tool and Figure 24 exemplifies the results provided by it.

The results of this tool – for instance, the result obtained for a female, 55 years old, with clinical tumor stage of 3, radiotherapy dose of 45 Gy, no concomittant chemotherapy, low anterior resection as surgery procedure, no adjuvant chemotherapy, 3 for pathological tumor stage and 1 for pathological nodal stage –, can be interpreted as would be if a group of 100 patients with the same characteristics as this individual patient, 67 patients would have no local recurrence, 50 patients would have no distant metastases, 63 patients would be alive 5 years after the treatment. Due to the fact that a model can never be completely the same as the "real world", these numbers could be lower or higher, but these are the most likely values. This particular patient has a high risk of developing a local recurrence,

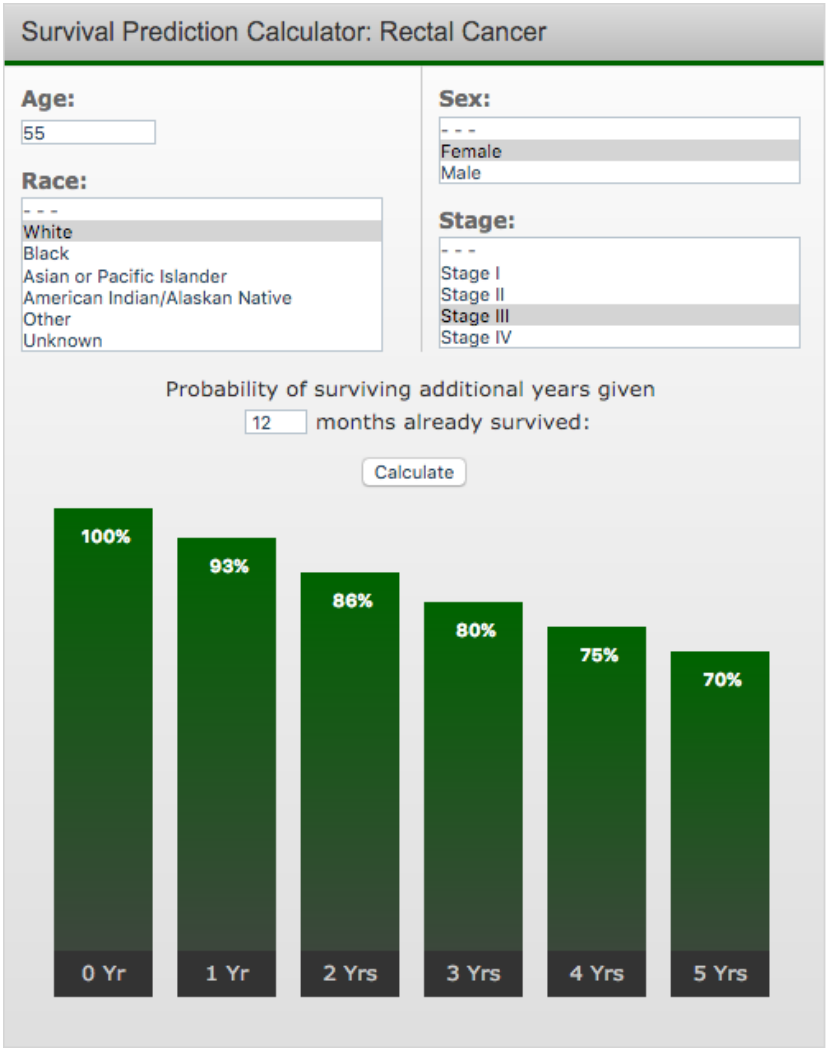



Figure 21.: Interactive tool for individualized estimation of conditional survival in rectal cancer and its results [101].

high risk of developing distant metastases and a high risk of dying within 5 years after treatment.

*Bowles et al. (2013)*

From The University of Texas M. D. Anderson Cancer Center, Bowles et al. [10] created an internet-based individualized conditional survival calculator[11] for patients with RC.

Taking into account the simultaneous effect of multiple variables on survival, separate multivariate Cox regression models were built for: no radiotherapy, preoperative radiotherapy, postoperative radiotherapy and stage IV patients. These models were created to

11 This prediction tool can be accessed at www.mdanderson.org/rectalcalculator.

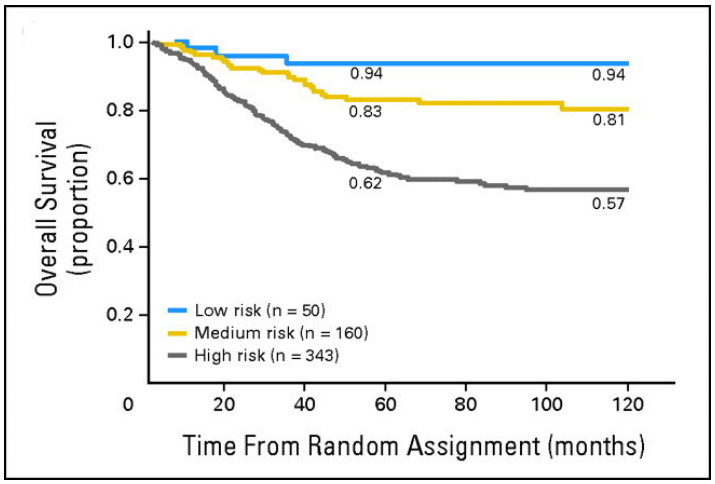



Figure 22.: Kaplan-Meier curves of risk group stratification for overall survival for validation dataset [98].

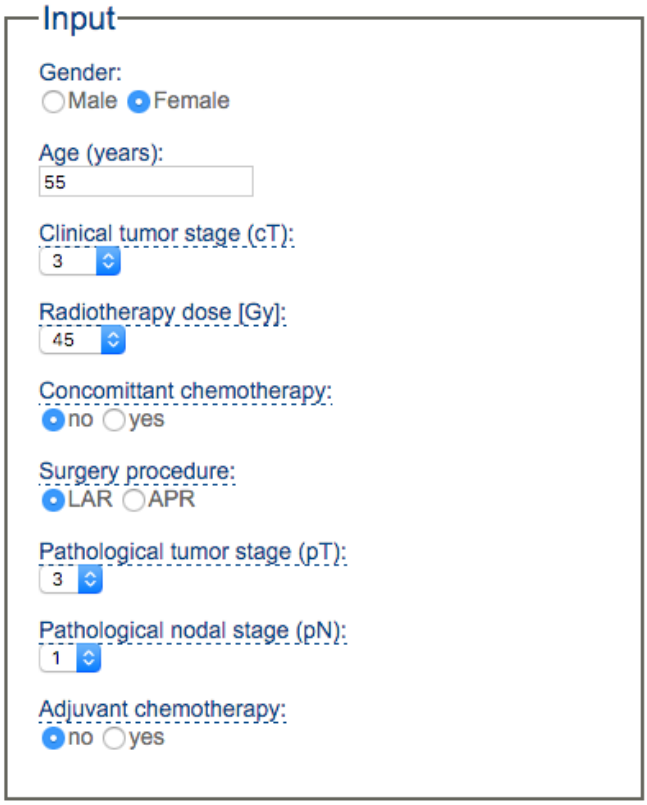



Figure 23.: Nomograms for predicting local recurrence, distant metastases, and overall survival for patients with locally advanced rectal cancer [97].

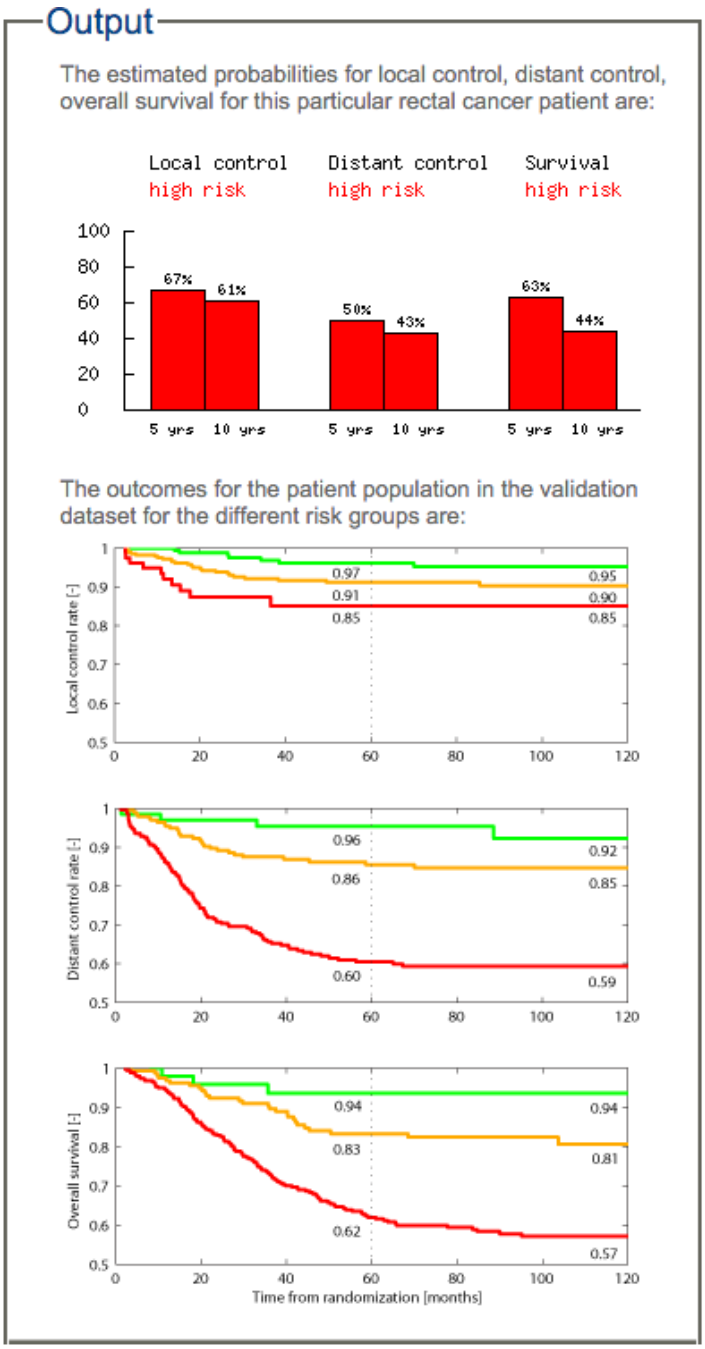



Figure 24.: Results of the nomograms for predicting local recurrence, distant metastases, and overall survival for patients with locally advanced rectal cancer [97].

determine adjusted survival estimates (at year 1 through 10) and used to calculate 5-year adjusted conditional survival. They were constructed using registries of 22,610 patients with rectal adenocarcinoma, who were diagnosed from January 1988 to December 2002, from the SEER database.

Models developed for patients with localized stage (stage I-III), i.e., for patients who underwent no radiotherapy, preoperative radiotherapy or postoperative radiotherapy, covariates were the same. They included age (<50, 50-59, 60-69, 70-79, 80+), sex (male, female), race (white, black, other), tumor grade (low [well-differentiated, moderately-differentiated], high [poorly-differentiated or undifferentiated] or unknown), surgery type (local excision and radical surgery) and AJCC sixth edition stage. In the model built for stage IV patients, i.e., for patients with distant metastasis, the surgery type was treated as a binary variable in the model (using any radiotherapy or primary tumor directed surgery as covariates).

The measures of performance for this tool are not available. Figure 25 shows the interface of this tool and Figure 26 its results for no radiotherapy. The other models are available clicking on the links shown in the figure.

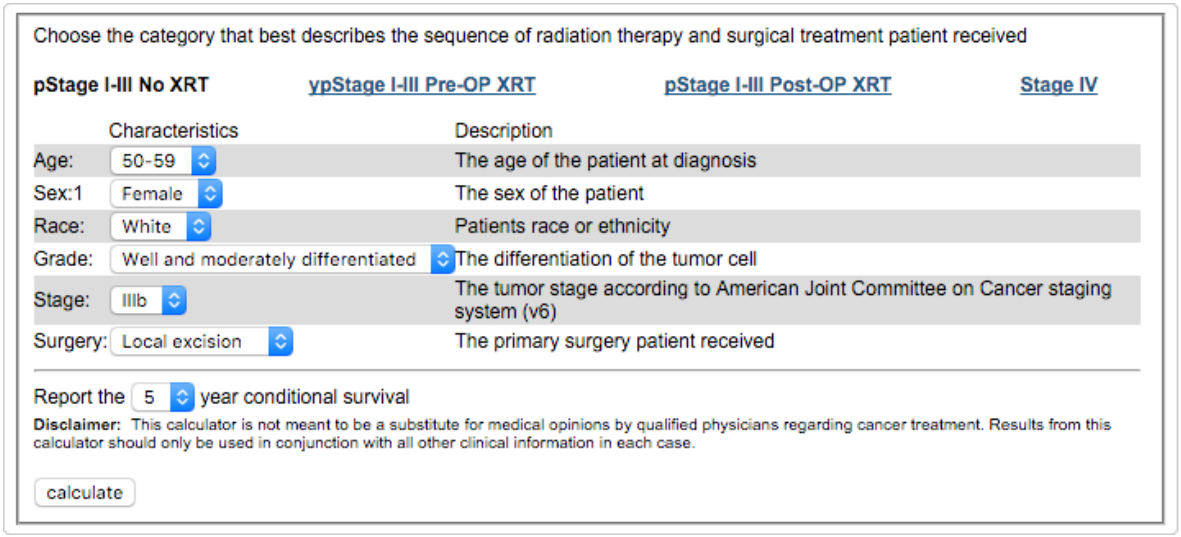



Figure 25.: Browser-based calculator to predict individualized disease-specific survival and conditional survival for rectal cancer patients [9].

## 2.2 PREDICTION MODELS

*Snow et al. (2001)*

Snow et al. [89] developed an *Artificial Neural Network (ANN)* model and a regression-based model to predict individual patient survival status 5 years after treatment.

Both models were developed employing 37,500 registries of colon carcinoma patients, from the *National Cancer Data Base (NCDB)*, United Kingdom. The data used were collected from 1985 to 1993.

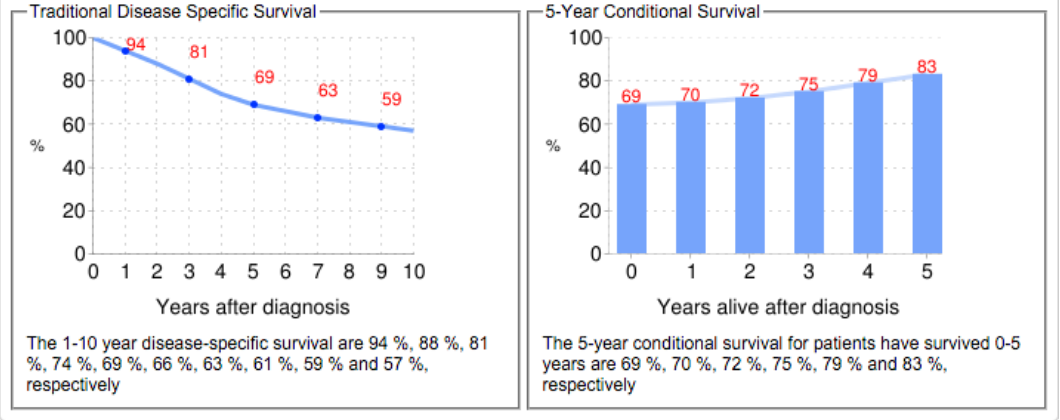



Figure 26.: Results of the browser-based calculator to predict individualized disease-specific survival and conditional survival for rectal cancer patients [9].

ANNs were applied to select the more important variables from the NCDB. This method was chosen in virtue of the ANN abilities to find patterns in complex data, with many variables. The logistic regression was used because of its widespread acceptance by biostatistians as a standard for prediction. Both methods were compared on a prospective set of patients that were not included in model development.

A sensitivity analysis method was used and the variables that resulted in significant loss of accuracy were dropped. As a result, the gender, age, number of positive regional lymph nodes, number of regional lymph nodes examined, pathologic T, N and M code of TNM, pathologic AJCC stage group, residual tumor (none, microscopic or macroscopic), if surgery was performed and if radiation therapy was performed were the selected variables to incorporate in the model. Figure 27 shows the ANN used in this analysis.

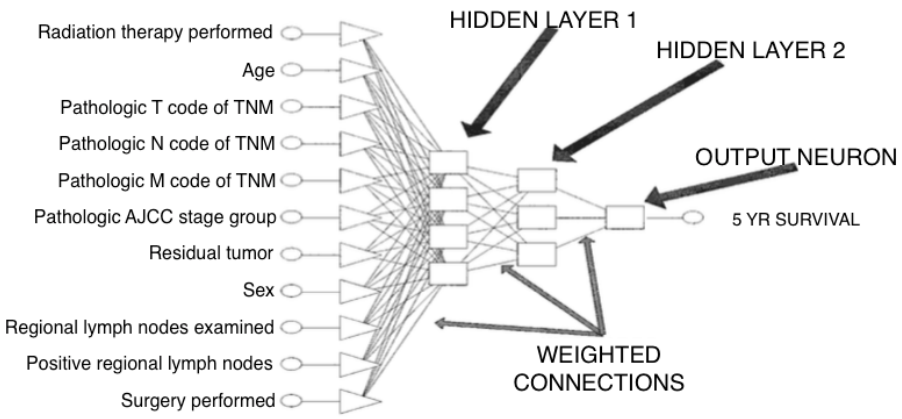



Figure 27.: ANN used in the analysis [89].

Figure 28 shows the survival stratified by pathologic Dukes stage in the NCDB database. The survival plot is a Kaplan–Meier type and uses a Cox model.

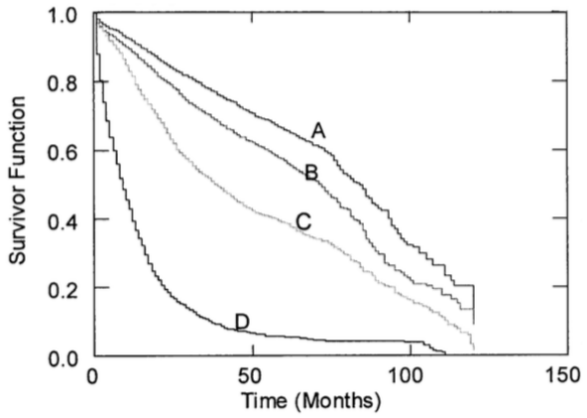



Figure 28.: Survival plot stratified by pathologic stage. Letters A, B, C, and D refer to Dukes Stages [89].

The area under the ROC curve was used to measure the overall predictive accuracy of the network. The ANN yielded a ROC area of 87.6%. A sensitivity to mortality of 95%, the specificity was 41%. The logistic regression provided a ROC area of 82%, and sensitivity to mortality of 95% gave a specificity of only 27%.

*Stojadinovic et al. (2012)*

Stojadinovic et al. [92] created a clinical decision support system using a *Machine-Learned Bayesian Belief Network (ML-BBN)* model for real-time estimation of overall survival in CC, providing personalized estimates of survival among patients.

A ML-BBN is a hierarchical network of associations between clinical factors in a registry data set that supplies multivariate mapping of complex data, allowing users to understand how different features are conditionally independent of each other [92, 91]. ML technology was employed because of its capability to capture complex, nonlinear, and in some cases, non-obvious patterns in a very large and heterogeneous data set.

The ML-BBN was constructed based on data from 146,248 records of patients with CC diagnosed between 1969 and 2006, from the SEER registry. From each registry independent prognostic factors were analyzed, including age and race of patient, the primary histology, grade and location of tumor. The number of primaries, AJCC T stage, N stage, and M stage were also considered.

Survival cohorts were developed based on follow-up time and overall survival time. To evaluate overall survival at different clinically relevant time points 4 subsets were created based on follow-up times of 12, 24, 36, and 60 months.

Models were compared with one another and with the AJCC TNM system by calculating a C-index, performing calibration, and identifying the area under ROC curves. Table 4 details these results. To validate the models, a Kaplan-Meier analysis was performed using the estimated mortality probabilities produced by the BBN as an additional validation method.

Table 4.: Comparative performance statistics – AJCC TNM Staging (Sixth Edition) vs. ML-BBN [92].

| Mortality | AUC | | PPV | | NPV | | Sensitivity | | Specificity | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AJCC | BBN | AJCC | BBN | AJCC | BBN | AJCC | BBN | AJCC | BBN |
| 12 months | 0.75 | 0.85 | 36.2 % | 74.4 % | 88.7 % | 85.1 % | 36.2 % | 51.4 % | 88.6 % | 94.0 % |
| 24 months | 0.76 | 0.85 | 54.6 % | 79.9 % | 81.1 % | 79.7 % | 54.6 % | 62.7 % | 81.1 % | 90.3 % |
| 36 months | 0.77 | 0.85 | 67.7 % | 81.8 % | 72.2 % | 73.9 % | 67.7 % | 69.9 % | 72.2 % | 84.5 % |
| 60 months | 0.77 | 0.85 | 85.9 % | 84.2 % | 47.7 % | 64.8 % | 85.9 % | 88.7 % | 47.7 % | 55.5 % |

The results showed that when compared with the AJCC staging system alone, these ML-BBN models showed superior sensitivity and specificity in estimating mortality. The larger area under the ROC curves (0.85) of the models shows that the ML-BBNs have a better discriminatory capacity in estimating survival within a defined period following initial cancer treatment. The large areas under the ROC curves were further confirmed using Kaplan-Meier (log rank) analysis that shows high, statistically significant odds ratios.

*Al-bahrani et al. (2013)*

Al-bahrani et al. [3] developed a survival prediction model for colon cancer, using ensemble data mining. In this work, supervised classification methods were used to predict survival of patients, at the end of 1 year, 2 years and 5 years of diagnosis.

The SEER data from 1973 to 2009 was analyzed and passed for a cleanup process, in a total of 105,133 registries. The original dataset with 134 attributes was reduced to 65, by removing useless attributes. Three new classes were created for 1 year, 2 years and 5 years survivability. This distribution of the data is shown in Table 5. After the cleanup process, there were a total of 65 attributes plus the class.

Table 5.: Class distribution of data [3].

| | Survival Classes | | |
|---|---|---|---|
| | 1 Year | 2 Year | 5 Year |
| Not Survived | 21.44% | 30.44% | 42.06% |
| Survived | 78.56% | 69.56% | 57.94% |

From the 65 attributes, a selection of attributes was performed using *Correlation Feature Selection (CFS)* [36] and *Information Gain Ratio (IGR)*, yielding 13 attributes. These selected attributes and their description can be observed in Table 6.

Table 6.: Selected Attributes.

| Attribute | Description |
|---|---|
| EOD-Extension | Documented tumor away from the primary site |
| SEER modified AJCC Stage 3rd ed (1988-2003) | The modified version stages cases that would be unstaged under strict AJCC staging rules |
| Birth Place | Place of birth encoded |
| EOD-Lymph Node Involv | Highest specific lymph node chain that is involved by the tumor |
| Regional Nodes Positive | Number of regional lymph nodes examined |
| RX Summ-Surg Prim Site | Surgical procedure for remove and/or destroy tissue |
| Histologic Type ICD-O-3 | Microscopic composition of cells and/or tissue for a specific primary |
| Reason for no surgery | Reason that surgery was not performed on the primary site |
| Age at diagnosis | Age of the patient at diagnosis |
| Diagnostic Confirmation | The best method used to confirm the presence of the cancer |
| EOD-Tumor Size | Largest dimension of the primary tumor (millimeters) |
| Behavior (92-00) ICD-O-2 | Behavior codes of the cancer |
| Primary Site | The site in which the primary tumor was originated |

The approach compares a model using the 65 attributes, acquired after filtering the data and removing the useless attributes, with another model using the 13 selected attributes which were obtained after running feature selection methods. As shown in Table 5, the dataset was imbalanced. Consequently, in the 13 selected attributes the *Synthetic Minority Over-sampling Technique (SMOTE)* was applied to generate synthetic examples by oversampling the minority class and introducing new synthetic patient records. The data that resulted of this process were used to construct another model, which was also compared with the previous developed models.

The WEKA toolkit was used to construct the models for survival prediction for colon cancer patients. Different basic and meta classification schemes were tested. The basic classification algorithms used were the J48 decision tree [69], REPTree [105], Random Forest [12], ADTree [30] and Logistic Regression [33]. In order to boost the basic classifiers and improve their performance, the following meta classifiers were used: Bagging [12], AdaBoost [32], Random Subspace [39] and Voting[12] [46].

The ensemble voting, composed by the top 3 performing classification schemes, was the best model. It had a predictive percentage accuracy of 90.38%, 88.01%, and 85.13% for 1 year, 2 years, and 5 years respectively and an area under the ROC curve of 0.96, 0.95, and 0.92 for 1 year, 2 years, and 5 years respectively.

12 Voting is a popular ensemble technique for combining multiple classifiers.

## 2.3 DISCUSSION

Throughout the chapter of the state of the art, several models to predict the survival of colorectal cancer patients were found. Some of them are available to physicians and patients in a web-based tool.

Table 7 shows the variables used in the applications and models to calculate the survival of CC patients.

Table 7.: Variables used in the applications and models for colon cancer patients.

| | Applications | | | | Models | | |
|---|---|---|---|---|---|---|---|
| | Bush and Michaelson (2009) [14] | Chang et al. (2009) [18] | Weiser et al. (2011) [104] | Renfro et al. (2014) [86] | Snow et al. (2001) [89] | Stojadinovic et al. (2012) [92] | Al-bahrani et al. (2013) [3] |
| Age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Gender | ✓ | ✓ | ✓ | ✓ | ✓ | × | × |
| Grade | ✓ | ✓ | ✓ | ✓ | × | ✓ | × |
| Stage | × | grouped stage | T- and N-stage | T-stage | TNM and grouped stage | TNM | grouped stage |
| CEA status | ✓ | × | × | × | × | × | × |
| Race | × | ✓ | × | ✓ | × | ✓ | × |
| Number of positive nodes | ✓ | × | ✓ | ✓ | ✓ | × | ✓ |
| Total number of nodes | × | × | ✓ | ✓ | ✓ | × | × |
| Tumor diameter | ✓ | × | × | × | × | × | ✓ |
| Histological type | ✓ | × | × | × | × | ✓ | ✓ |
| Site of the colon | ✓ | × | × | ✓ | × | ✓ | ✓ |
| Surgery | × | × | × | × | ✓ | × | ✓ |
| Radiation therapy | × | × | × | × | ✓ | × | × |
| Residual tumor | × | × | × | × | ✓ | × | × |
| Number of primaries | × | × | × | ✓ | × | ✓ | × |
| Extension of tumor | ✓ | × | × | × | × | × | ✓ |
| Birth place | × | × | × | × | × | × | ✓ |
| Lymph node involved† | × | × | × | × | × | × | ✓ |
| Reason for no surgery | × | × | × | × | × | × | ✓ |
| Diagnostic confirmation | × | × | × | × | × | × | ✓ |
| Behavior | × | × | × | × | × | × | ✓ |
| Body mass index | × | × | × | ✓ | × | × | × |
| Treatment type | × | × | × | ✓ | × | × | × |
| Performance status | × | × | × | ✓ | × | × | × |

† Contained in TNM stage.

Table 8 shows the overall characteristics of models (with and without application available to users) for CC patients.

Table 8.: Characteristics of models (with and without an application available to users) for colon cancer patients.

| | Applications | | | | Models | | |
|---|---|---|---|---|---|---|---|
| | Bush and Michaelson (2009) [14] | Chang et al. (2009) [18] | Weiser et al. (2011) [104] | Renfro et al. (2014) [86] | Snow et al. (2001) [89] | Stojadinovic et al. (2012) [92] | Al-bahrani et al. (2013) [3] |
| Number of Variables | 9 | 6‡ | 2/3/7 | 12 | 9 | 7 | 13 |
| Dataset | SEER | SEER | SEER | ACCENT | NCDB | SEER | SEER |
| Model | regression-based | regression-based | regression-based | regression-based | ML-based and regression-based | ML-based | ML-based |
| Target | 0 – 15 years | 1 – 10 years (disease specific survival) 0 – 5 years (conditional survival) | 5 years | 5 years | 5 years | 1, 2, 3 and 5 years | 1, 2 and 5 years |
| Performance | – | C-index: 0.816 | C-index: 0.61/0.63/0.68 | C-index: 0.66 | AUC: 0.876 (ANN) AUC: 0.82 (logistic regression) | AUC: 0.85 | AUC: 0.96/0.95/0.92 Accuracy: 90.38%/88.01%/85.13% |

‡ Including months which already survived (for conditional survival calculate).

In a review of literature, none of models without an application for RC patients was found. Table 9 shows the variables used in the tools for rectal cancer patients.

Table 9.: Variables used in the applications for rectal cancer patients.

| | Applications | | |
|---|---|---|---|
| | Wang et al. (2011) [102] | Valentini et al. (2011) [98] | Bowles et al. (2013) [10] |
| Age | ✓ | ✓ | ✓ |
| Gender | ✓ | ✓ | ✓ |
| Grade | × | × | ✓ |
| Stage | grouped stage | clinical T-, pathological T- and pathological N-stage | grouped stage |
| Race | ✓ | × | ✓ |
| Surgery Procedure | × | ✓ | ✓ |
| Radiotherapy dose | × | ✓ | × |
| Concomittant chemotherapy | × | ✓ | × |
| Adjuvant chemotherapy | × | ✓ | × |

Table 10 shows the overall characteristics of models with an application available to users for RC patients.

Table 10.: Characteristics of models (with an application available to users) for rectal cancer patients.

| | Wang et al. (2011) [102] | Valentini et al. (2011) [98] | Bowles et al. (2013) [10] |
|---|---|---|---|
| Number of Variables | 5‡ | 9 | 7‡ |
| Dataset | SEER | five European randomized trials | SEER |
| Model | regression-based | regression-based | regression-based |
| Target | 0 – 5 years | 1– 10 years | 1 – 10 years (disease specific survival)<br>0 – 5 years (conditional survival) |
| Performance | C-index: 0.75 | C-index: 0.70 | – |

‡ Including months which already survived (for conditional survival calculate).

Observing the variables used to calculate the survival of patients with colon cancer, only the age of the patient is common to all models (with and without an application available). The stage of cancer (in grouped form), gender of patient, grade of tumor and the number of positive nodes are the other common variables. All the tools used to predict the survival of rectal cancer patients are to calculate the conditional survival and all of them employ the age and gender of patients. The stage of cancer (in grouped form), race of patient and surgery procedure are the other common variables. Moreover, the way in which the variables were selected to be part of model was not always evident. The present work intends to use data mining techniques to select the most relevant features and compare them to the opinion of a specialist physician, using the prognostic factors and the relations between them. Another goal is to know if similar features are selected for both cancers.

In order to assess the performance of models, all works described herein applied one of two metrics: the C-index and the *The Area Under the ROC Curve (AUC)*. These measures are considered to be numerically identical [38]. They correspond to the probability of giving a correct response in a binary prediction problem. A value of 1 means a perfect model, whereas a value of 0.5 indicates a random guessing model [43, 44, 54]. According to this, the model developed by Al-bahrani et al. [3] is the work that had the best perfomance values for colon cancer survival prediction (with an AUC of 0.96, 0.95 and 0.92, for 1, 2 and 5year, respectivelly). However, this work is not available under any platform to health care professionals. For rectal cancer survival prediction, the best results known belong to Wang et al. [102] (with a C-index of 0.75).

All the presented tools utilize regression models. Ahmed et al. [2] compared a regression-based model with a ML-based model, having obtained better results of performance using ML. Modeling with ML techniques allows to find underlying patterns and makes possible to deal with missing information. In this work we intend to develop a model using machine learning techniques.

To determine if related tools are suitable to mobile devices, all applications (for both cancers) were analyzed using the mobile-friendly test tool of *Google*[13]. The tool of *Google* reported all applications, except the ones developed for colon cancer patients by Weiser et al. [104] and Renfro et al. [86], are unsuitable to access via smartphone or even tablet. The test revealed that the text was too small to read, the mobile viewport was not set, links were too close together or content was wider than the screen. Therefore, none of these applications had a mobile-friendly design. Another goal is to address this and develop a cross-platform tool, that is available to users in a practical and intuitive way, through a smartphone or tablet.

13 Mobile-friendly test tool of *Google* is available at https://www.google.com/webmasters/tools/mobile-friendly/

# 3

# DEVELOPMENT OF THE PREDICTION MODEL

This chapter is the main of the dissertation. It describes all the processes of the development of the prediction model, from the raw data to the modeling and evaluation phase, including testing. Prediction models were developed to calculate the survival and conditional survival of colon and rectal patients after diagnosis and treatment, taking into account the cover a 5-year span – an important goal for a colorectal cancer patient to overcome. The conditional survival model was developed applying the methods which produced the best results for the the survival models.

## 3.1 RAW DATA IMPORTING

The survival prediction systems, for colon and rectal cancer, were projected to produce an individualized response. Each system was planned to receive a number of inputs for selected prediction features and, for each of the 5 years following treatment, generate an output stating whether the patient in question will survive that year or not, along with a confidence value for the prediction. In case of the conditional survival prediction systems, the process is similar. Depending on the years that the patient has already survived, the generated output is given for each following years, until the fifth year after treatment.

The National Cancer Institute provides the access to the largest population-based cancer database in the United States of America, by the SEER Program. This database contains 8,689,771 cases collected from 1973 to 2012, including several types of cancer. Its registries are available in the binary format. For that reason, it was required create a script to convert the relevant dataset into an intelligible form, for later be imported to a data mining software (Figure 29). The software chosen to develop the prediction model was RapidMiner Studio (6.5 version)[1], an open source data mining software. It has a workflow-based interface that allows an clear construction of complex data management processes. Moreover, it offers an intuitive *Application Programming Interface (API)*.

1 Software available at http://rapidminer.com/.

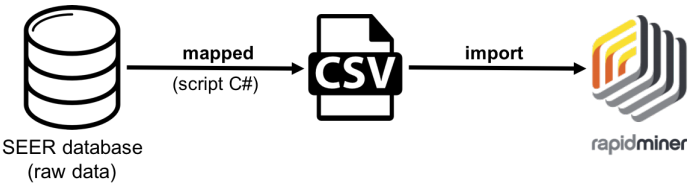



Figure 29.: From raw data to RapidMiner Studio software.

The development of a prediction model requires several phases, from the preprocessing of SEER data to the selection of the best model. All of the phases are described in the ensuing sections. It is important to clarify that the survival prediction was handled as a classification problem and five classification models, for each year and type of cancer (colon and rectal), were developed. Theses models were posteriorly combined, in a programmatic manner, into a model capable of providing a prediction for each year with a single interaction, within the selected cancer type.

## 3.2 PREPROCESSING

The colorectal cancer data from SEER contained 515,791 records, from 1973 to 2012, and consisted of 146 attributes, some of them only applicable to a limited period within the time of data collection.

During the Preprocessing phase, it was defined that the period of interest would be from 2004 onwards, minimizing the occurrence of missing data due to the applicability of the attributes. This operation was determined filtering [75] the data by the year of diagnosis.

Additionally, empty attributes, attributes that are not applicable to this type of cancer (e.g., the human epidermal growth factor receptor 2 result, an indicator used in breast cancer only [106]) and attributes that are not directly related with the vital status of the patient were removed (e.g. the number identifying the registry of the patient). It was defined by using the Select Attributes operator [83] and selecting the pertinent attributes.

Only the adult patients (age greater than or equal to 18 years old) were selected for further processing (filtering [75] by the age of diagnosis) and the missing values were replaced (using the Replace Missing Values operator [80] of RapidMiner software) applying the *unknown* code.

Patients who were alive at the end of the data collection whose survival time had not yet reached 60 months (five years), the maximum period for which the model under development is supposed to predict survival, and those who passed away of causes other than colon or rectal cancer were sampled out (filtering by survival months and the cause of death to SEER site [75]) from the training set as their inclusion was considered to be unsuited to the problem at hand. The numeric attributes were converted to nominal [77] (e.g. sex) and

the binary classes (*survived* and *not survived*) were derived for the target labels 1-, 2-, 3-, 4- and 5-year survival.

Finally, based on existing attributes and at the request of a physician who collaborated in this work, new attributes, such as the number of regional lymph negative nodes, the ratio of positive nodes over the total examined nodes and also the relapse of the patients for colon cancer, were calculated. It was defined using aggregate functions [72] and the Generate Attributes operator [76].

After the Preprocessing phase, the attributes were reduced to 61, including the new attributes and the target labels and the data was reduced to 51,410 records: 38,592 and 12,818 registries, for colon and rectal cancer, respectively. From the isolated cases for each pathology, 10% were randomly selected for a testing set and the remaining were used to developed the prediction models.

## 3.3 SPLIT DATASET

For the survival prediction, the second phase consisted in divide the data into five sub-datasets (for each cancer type). The data was split by target label, according to the corresponding survival year. Table 11 and Table 12 shows the class distribution in each sub-dataset, for colon and rectal cancer, respectively.

Table 11.: Class distribution for each target label in the sub-datasets, for survival models and colon cancer.

| | Target Labels | | | | |
|---|---|---|---|---|---|
| | 1 Year | 2 Year | 3 Year | 4 Year | 5 Year |
| Not Survived | 24.51% | 32.60% | 36.96% | 39.35% | 41.07% |
| Survived | 75.49% | 67.40% | 63.04% | 60.65% | 58.93% |
| Total number of cases | | | 34,732 | | |

Table 12.: Class distribution for each target label in the sub-datasets, for survival models and rectal cancer.

| | Target Labels | | | | |
|---|---|---|---|---|---|
| | 1 Year | 2 Year | 3 Year | 4 Year | 5 Year |
| Not Survived | 4.03% | 5.89% | 7.17% | 8.08% | 8.70% |
| Survived | 87.88% | 82.27% | 78.41% | 75.68% | 73.79% |
| Total number of cases | | | 11,536 | | |

For the conditional survival prediction, the data was split into ten datasets (for each type of cancer). The data were separated by target label, according to the corresponding survival year and the years that patients had already survived. For instance, taking into account that

a patient which already survived the first year after treatment, will be necessary calculate the outcome from the second to fifth year. In this way, registries from patients who died during the first year were not included of the sub-dataset. Table 13 and Table 14 shows the class distribution in each sub-dataset, for colon and rectal cancer, respectively.

Table 13.: Class distribution for each target label in the sub-datasets, for conditional survival models and colon cancer.

| | Target Labels | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | survived the 1st year | | | | survived the 2nd year | | | survived the 3rd year | | survived the 4th year |
| | 2 Year | 3 Year | 4 Year | 5 Year | 3 Year | 4 Year | 5 Year | 4 Year | 5 Year | 5 Year |
| Not Survived | 11.45% | 18.77% | 22.30% | 24.40% | 8.15% | 12.15% | 14.55% | 4.24% | 6.84% | 2.77% |
| Survived | 88.55% | 81.23% | 77.70% | 75.60% | 91.85% | 87.85% | 85.45% | 95.76% | 93.16% | 97.23% |
| Total number of cases | 15,765 | | | | 13,969 | | | 12,816 | | 12,261 |

Table 14.: Class distribution for each target label in the sub-datasets, for conditional survival models and rectal cancer.

| | Target Labels | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | survived the 1st year | | | | survived the 2nd year | | | survived the 3rd year | | survived the 4th year |
| | 2 Year | 3 Year | 4 Year | 5 Year | 3 Year | 4 Year | 5 Year | 4 Year | 5 Year | 5 Year |
| Not Survived | 93.71% | 89.10% | 86.50% | 84.26% | 95.22% | 92.58% | 90.14% | 97.18% | 94.72% | 97.49% |
| Survived | 6.29% | 10.90% | 13.50% | 15.74% | 4.78% | 7.42% | 9.86% | 2.82% | 5.28% | 2.51% |
| Total number of cases | 4,421 | | | | 4,139 | | | 3,939 | | 3,826 |

## 3.4 FEATURE SELECTION

The Feature Selection phase was crucial to determine the most influential features on the survival of colon and rectal cancer patients. In order to accomplish this the Optimize Selection operator [79] of RapidMiner was used. It implements a deterministic and optimized selection process with decision trees and *forward selection*. The process was applied to each sub-dataset for the target label. Only the features in common to all the sub-datasets, for each cancer, were selected and used to construct the prediction models. Table 15 and Table 16 shows the selected features for colon and rectal cancer, respectively. It also shows the meaning of the selected features.

The 6 selected features, of each cancer, were compared with a set of 18 features (shown in Table 17) indicated by a specialist physician on colorectal cancer. These indicated attributes were given, as common, for both types of cancer. The three sets of features were mapped to attributes in the sub-datasets and later used to generate and evaluate the prediction models.

Table 15.: Attributes selected in the Feature Selection process for CC.

| Attribute | Description |
|---|---|
| Age recode with < 1 year old | Age groupings based on age at diagnosis (single-year ages) of patients (< 1 year, 1-4 years, 5-9 years, ..., 85+ years) |
| CS Site-Specific Factor 1 | The interpretation of the highest Carcinoembryonic Antigen (CEA) test results |
| CS Site-Specific Factor 2 | The clinical assessment of regional lymph nodes |
| Derived AJCC Stage Group | The grouping of the TNM information combined |
| Primary Site | Identification of the site in which the primary tumor originated |
| Regional Nodes Examined | The total number of regional lymph nodes that were removed and examined by the pathologist |

Table 16.: Attributes selected in the Feature Selection process for RC.

| Attribute | Description |
|---|---|
| Age recode with < 1 year old | * |
| CS Extension | Extension of the tumor |
| CS Tumor Size | Size of the tumor (in mm) |
| Derived AJCC Stage Group | * |
| RX Summ–Surg Prim Site | Describes a surgical procedure that removes and/or destroys tissue of the primary site performed as part of the initial work-up or first course of therapy. |
| Sex | The gender of the patient at diagnosis |

* Described in Table 15.

Table 17.: Attributes selected by a specialist physician on CC.

| Attribute | Description |
|---|---|
| Age at Diagnosis | The age of the patient at diagnosis (continuous value) |
| CS Extension | ◇ |
| CS Site-Specific Factor 8 | The perineural Invasion |
| CS Tumor Size | ◇ |
| Derived AJCC T, N and M | The AJCC T, N and M stage (6th ed.) |
| Grade | Grading and differentiation codes |
| Histologic Type | The microscopic composition of cells and/or tissue for a specific primary |
| Laterality | The side of a paired organ or side of the body on which the reportable tumor originated |
| Primary Site | * |
| Race Recode (White, Black, Other) | Race recode based on the race variables |
| Regional Nodes Examined | * |
| Regional Nodes Positive | The exact number of regional lymph nodes examined by the pathologist that were found to contain metastases |
| Regional Nodes Negative | (Regional nodes examined - Regional nodes positive) |
| Regional Nodes Ratio | (Regional nodes negative over Regional nodes examined) |
| Relapse | The relapse of the patients for colon cancer |
| Sex | ◇ |

* Described in Table 15.
◇ Described in Table 16.

## 3.5 DATA SAMPLING

As observed, in Table 11 and Table 12, the classes are not equally represented. In order to determine if unbalanced datasets are or not a problem, these were compared with balanced data-sets. These balanced datasets were generated through the oversampling of the minority class, undersampling of the majority class and hybrid sampling (doing oversampling of the minority class and undersampling of the majority class). Also was determined the influence of the unknown values, in relation to the selected attributes. Using the Sample (Bootstrapping) operator [82], the minority class (the "not survived" class) of each year was oversampled to the corresponding "survived" value, for colon and rectal cancer. The majority class (the "survived" class) of each year was undersampled to the corresponding "not survived" value using the Sample operator [81], for each type of cancer. For the hybrid sampling, the minority class was oversampled and the majority class was undersampled for each year and cancer type, using the same operators aforementioned.

## 3.6 MODELING

The classification strategies used in the Modeling phase consisted of ensemble methods. The classification schemes applied were meta-classifiers. This type of classifier is used to boost basic classifiers and improve their performance. All the possible combinations of the classifiers were explored, according to the algorithms and type of attributes allowed. The tested meta-classifiers were:

- **Bagging** [12]: Also called bootstrap aggregating. It splits the data into $m$ different training sets on which $m$ classifiers are trained. The final prediction results from the equal voting of each generated model on the correct result. Bagging is used to improve stability and classification accuracy, reduce variance and avoid overfitting.
- **AdaBoost** [31]: This meta-classifier calls a new weak classifier at each iteration. A weight distribution which indicates the weight of examples in the classification is updated. It focuses on the examples that have been misclassified so far in order to adjust subsequent classifiers and reduce relative error.
- **Bayesian Boosting** [74]: A new classification model is produced at each iteration and the training set is reweighed so that previously discovered patterns are sampled out. The inner classifier is sequentially applied and the resulting models are later combined into a single model. The boosting operation is conducted based on probability estimates. It is particularly useful for discovering hidden groups in the data.
- **Stacking** [25]: This meta-classifier is used to combine base classifiers of different types. Each base classifier generates a model using the training set, then a meta-learner

integrates the independently learned base classifier models into a high level classifier by re-learning a meta-level training set. This meta-level training set is obtained by using the predictions of base classifiers in the validation dataset as attribute values and the true class as the target.

- **Voting** [46]: Each inner classifier of the meta-classifier receives the training set and generates a classification model. The prediction of an unknown example results from the majority voting of the derived classification models.

Since survival prediction is being handled as a classification problem, a group of basic classifiers were selected to be used in ensembles with the above-described meta-classifiers. The group includes some of the most widely used learners [78] available in RapidMiner. The tested basic classifiers were:

- **k-NN (Lazy Modeling)** [37]: this algorithm is based on learning by analogy. The training examples are described by $n$ attributes and each of them represents a point in a $n$-dimensional space. The test example is compared with them by searching the pattern space and it is classified according the $k$ training examples closest to it. The similarity is determined in terms of a distance metric, such as the Euclidean distance.

- **Naive Bayes (Bayesian Modeling)** [96]: it is a simple probabilistic classifier, based on the application of the Bayes theorem with strong (naive) assumption of independence between every pair of features.

- **Decision Tree (Tree Induction)** [70]: the data is classified using a hierarchical splitting mechanism (repeatedly splitting on the values of attributes), looking like an inverted tree with the root at the top and it growing downwards. Each node of tree corresponds to one of the input attributes. Normally, the recursion stops when all or the most of the examples or instances have the same label value.

- **Random Forest (Tree Induction)** [48]: is generated a set of a specified number of random trees, working like the Decision Tree. However, it uses only a random subset of attributes for each split. The resulting model is a voting model of all the random trees.

A total of fourteen classification schemes were explored for each type of cancer, set of attributes (6 and 18 attributes) and type of dataset (balanced or not) for 1, 2, 3, 4, and 5 survival years. The learning combinations of meta-classifiers with basic classifiers are as follows. The Stacking model used k-NN, Decision Tree, and Random Forest classifiers as base learners, and a Naive Bayes classifier as a Stacking model learner. The Voting model used k-NN, Decision Tree and Random Forest as base learners. The other models were used in combination with each basic classifier.

Table 18.: Table of confusion.

| | Labeled as Survivor | Labeled as Not Survivor |
|---|---|---|
| Predicted as Survivor | TP<br>True Positives | FP<br>False Positives |
| Predicted as Not Survivor | FN<br>False Negatives | TN<br>True Negatives |

## 3.7 EVALUATION

### 3.7.1 *Cross-validation*

For evaluation purposes, 10-fold cross-validation [84] was used to assess the prediction performance of the generated prediction models and avoid overfitting. In this process, the data is split into ten nearly identical portions, and each in turn is used for testing while the remnant is applied for training. The process is repeated ten times, in order that in the end every instance has been used exactly once for testing. The final validation result is the average of the 10 repetitions.

In classification problems, there are many ways to evaluate a classifier. A confusion matrix, also known as a contingency table, is usually employed to summarize the relationship between a classifier and an instance, in a binary or binomial classification [28]. In the context of this dissertation, a patient (the instance) can be classified as "survivor" or "not survivor" (Table 18) after 1, 2, 3, 4 or even 5 year after the diagnosis of colon or rectal cancer. When a patient is survivor, in a given year, and he is (well) classified as "survivor", it is counted as a true positive (TP). On the other hand, if he is classified as not survivor, it is counted as a false negative (FN). If the patient is not survivor and it is (well) classified as "not survivor", it is counted as a true negative (TN), but if he is classified as "survivor", it is counted as a false positive (FP).

Each classification scheme was evaluated using metrics based on confusion matrix: the prediction accuracy (Equation 4), the F-measure (Equation 5) and the AUC for 1, 2, 3, 4, and 5 years. The accuracy is the percentage of correct responses among the examined cases [11]. The F-measure is a combine of precision (a form of accuracy, also known as positive predictive value) and recall (also known as sensitivity) measures [68]. The AUC as it is an area, is calculated by an integral. Numerical methods, like the trapezoidal rule can be used to approximate the integral [27]. This measure can be interpreted as the percentage of randomly drawn data pairs of individuals that have been accurately classified in the two populations [47], and it is commonly used as a measure of quality for classification models [11].

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{4}$$

$$F - measure = 2\frac{(precision \times recall)}{(precision + recall)} = 2\frac{(\frac{TP}{(TP+FP)}) \times (\frac{TP}{(TP+FN)})}{(\frac{TP}{(TP+FP)} + \frac{TP}{(TP+FN)})} = \frac{2TP}{(2TP + FP + FN)} \tag{5}$$

### 3.7.2 *Testing*

Last but not least, the 10% separated data for testing were applying for each developed model, using the Apply Model operator [73]. A new attribute was generated to compare the predicted value with the real value of survival, in order to determine how well the model was able to receive new cases and give reliable predictions.

# 4

# EXPERIMENTAL RESULTS

In this chapter are presented the results of performance of each developed ensemble model. A total of 14 classification schemes were evaluated, for 6 and 18 attributes, type of dataset (balanced or not) and from 1 to 5 years after the diagnosis and treatment of patients with colon and rectal cancer. In order to realize the best classification scheme was calculated the average performance of each model. From the testing phase, the percentage of wrongly classified cases is presented for the same parameters that the models were evaluated. For the conditional survival models, they were developed taking into account the best results for the the survival models, thus, only a classification scheme was evaluated.

## 4.1 SURVIVABILITY PREDICTION MODELS

### 4.1.1 *Colon Cancer*

Figure 30 shows the average performances in terms of accuracy of the best learning schemes for the 5 years for the models trained with unbalanced datasets, unbalanced datasets without unknowns values, balanced oversampled datasets, balanced undersampled datasets and hybrid balanced datasets, for 6 and 18 attributes. The two best values for 18 and 6 attributes are labeled.

Figure 31 shows the average performances in terms of AUC of the best learning schemes for the 5 years for the models trained with unbalanced datasets, unbalanced datasets without unknowns values, balanced oversampled datasets, balanced undersampled datasets and hybrid balanced datasets, for 6 and 18 attributes.

Figure 32 shows the average performances in terms of F-measure of the best learning schemes for the 5 years for the models trained with unbalanced datasets, unbalanced datasets without unknowns values, balanced oversampled datasets, balanced undersampled datasets and hybrid balanced datasets, for 6 and 18 attributes.

Figure 33 shows the average percentage of wrongly classified cases of the best learning schemes for the 5 years for the models trained with unbalanced datasets, unbalanced

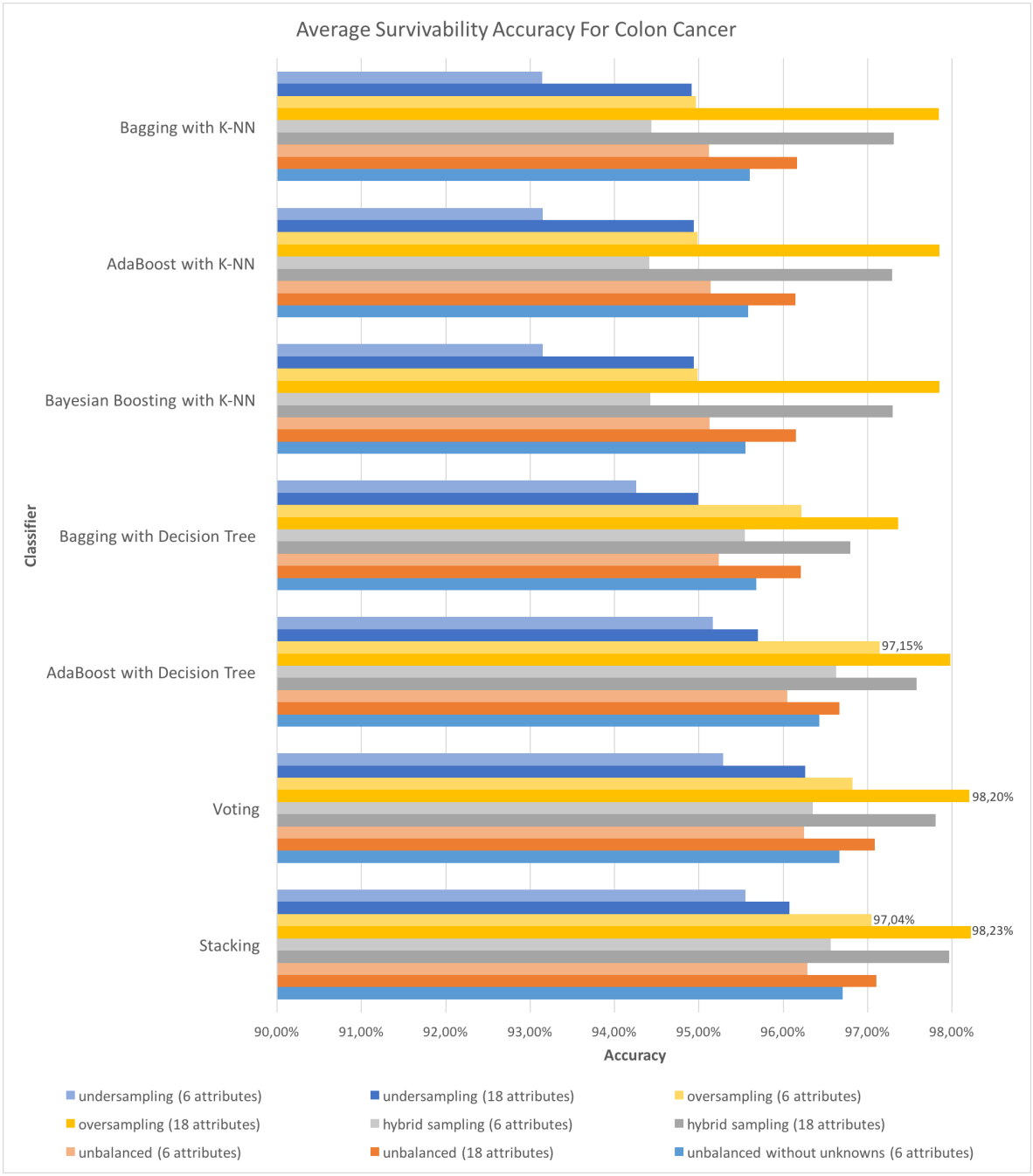



Figure 30.: Average survivability percentage accuracy for colon cancer: comparison of the 18-attribute models with the 6-attribute models for the best learning schemes.

datasets without unknowns values, balanced oversampled datasets, balanced undersampled datasets and hybrid balanced datasets, for 6 and 18 attributes.

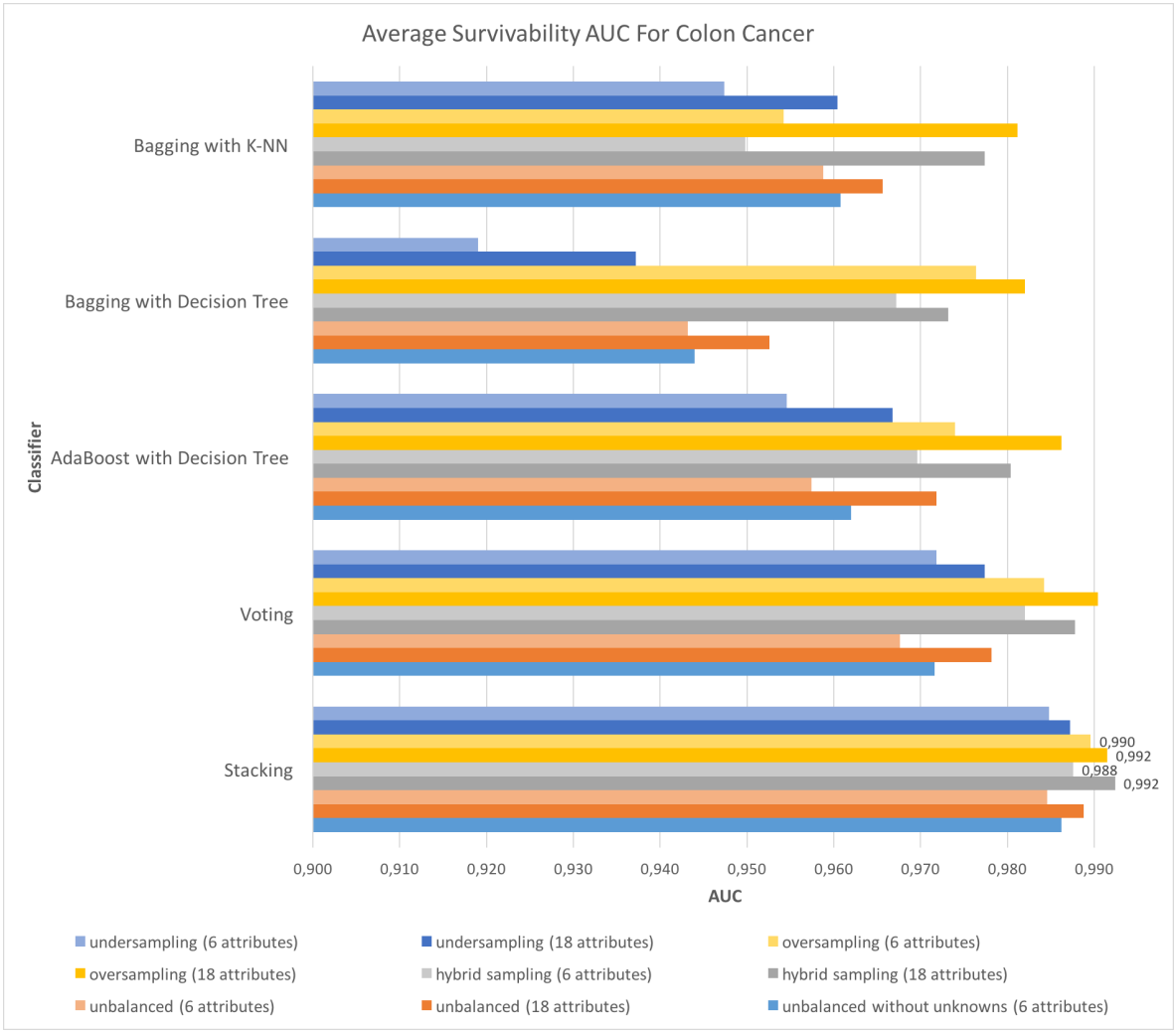



Figure 31.: Average survivability AUC for colon cancer: comparison of the 18-attribute models with the 6-attribute models for the best learning schemes.

### 4.1.2 *Rectal Cancer*

Figure 34 shows the average performances in terms of accuracy of the best learning schemes for the 5 years for the models trained with unbalanced datasets, unbalanced datasets without unknowns values, balanced oversampled datasets, balanced undersampled datasets and hybrid balanced datasets, for 6 and 18 attributes.

Figure 35 shows the average performances in terms of AUC of the best learning schemes for the 5 years for the models trained with unbalanced datasets, unbalanced datasets without unknowns values, balanced oversampled datasets, balanced undersampled datasets and hybrid balanced datasets, for 6 and 18 attributes.

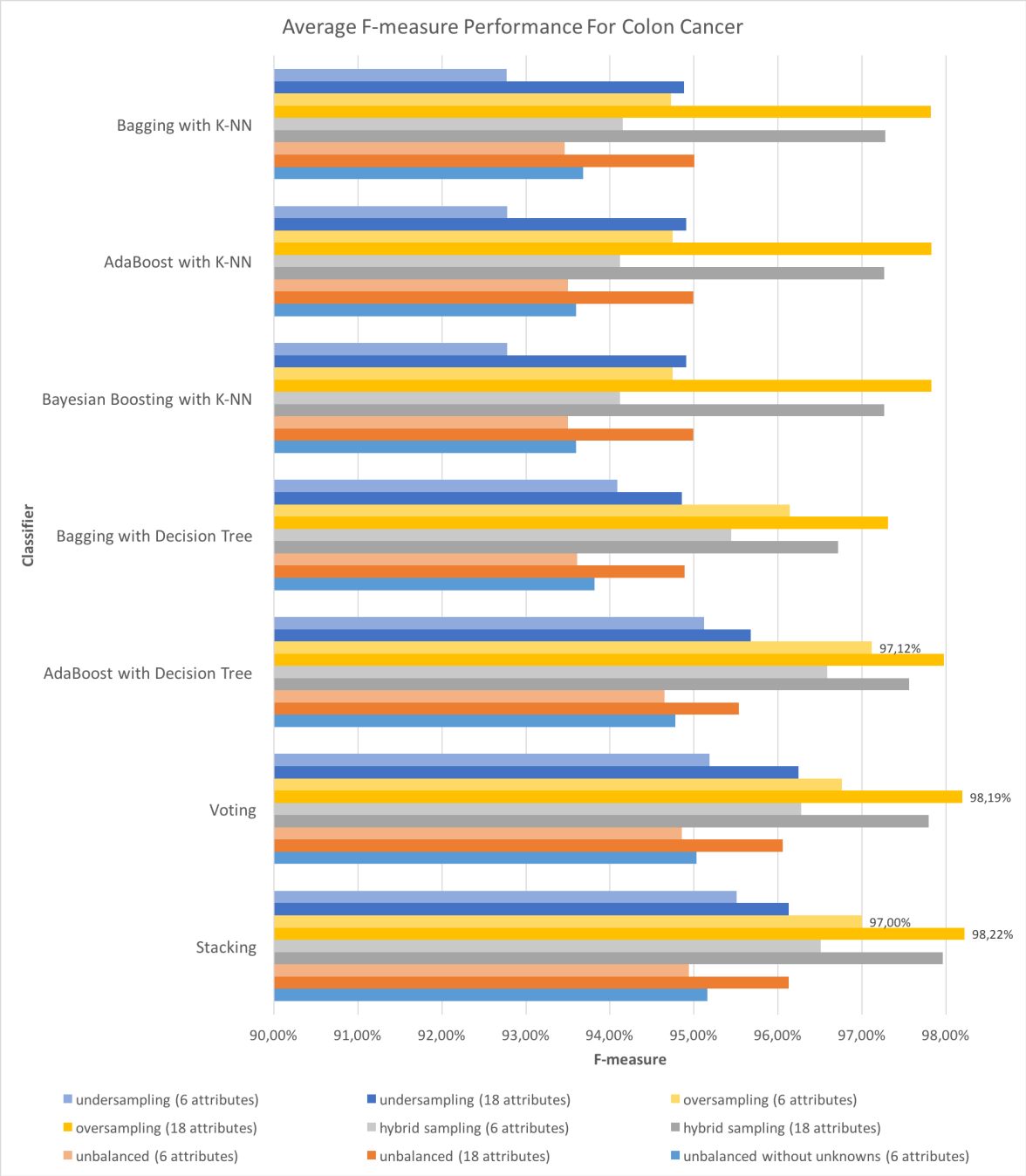



Figure 32.: Average F-measure performance for colon cancer: comparison of the 18-attribute models with the 6-attribute models for the best learning schemes.

Figure 36 shows the average performances in terms of F-measure of the best learning schemes for the 5 years for the models trained with unbalanced datasets, unbalanced datasets without unknowns values, balanced oversampled datasets, balanced undersampled datasets and hybrid balanced datasets, for 6 and 18 attributes.

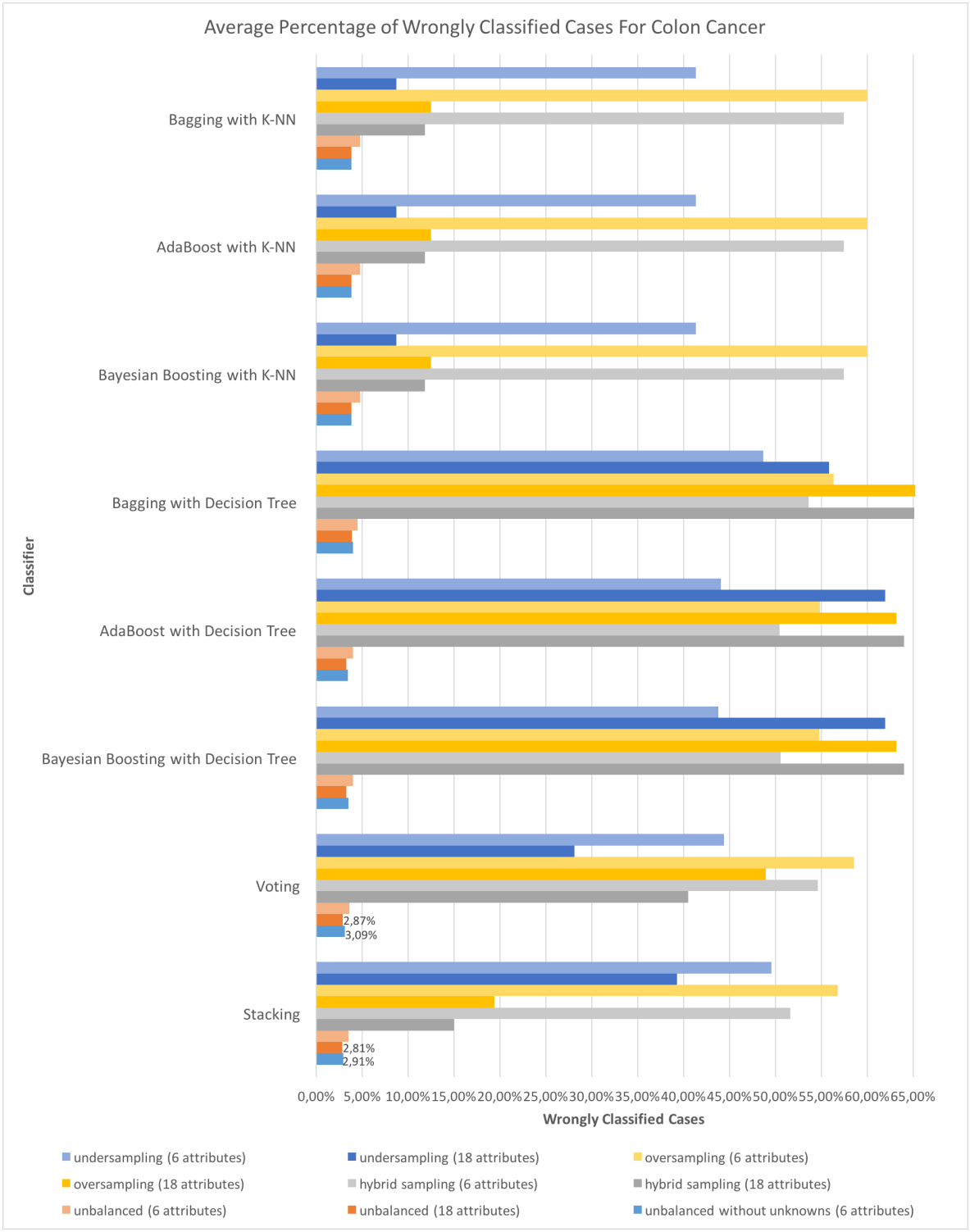



Figure 33.: Average percentage of wrongly classified cases for colon cancer: comparison of the 18-attribute models with the 6-attribute models for the best learning schemes.

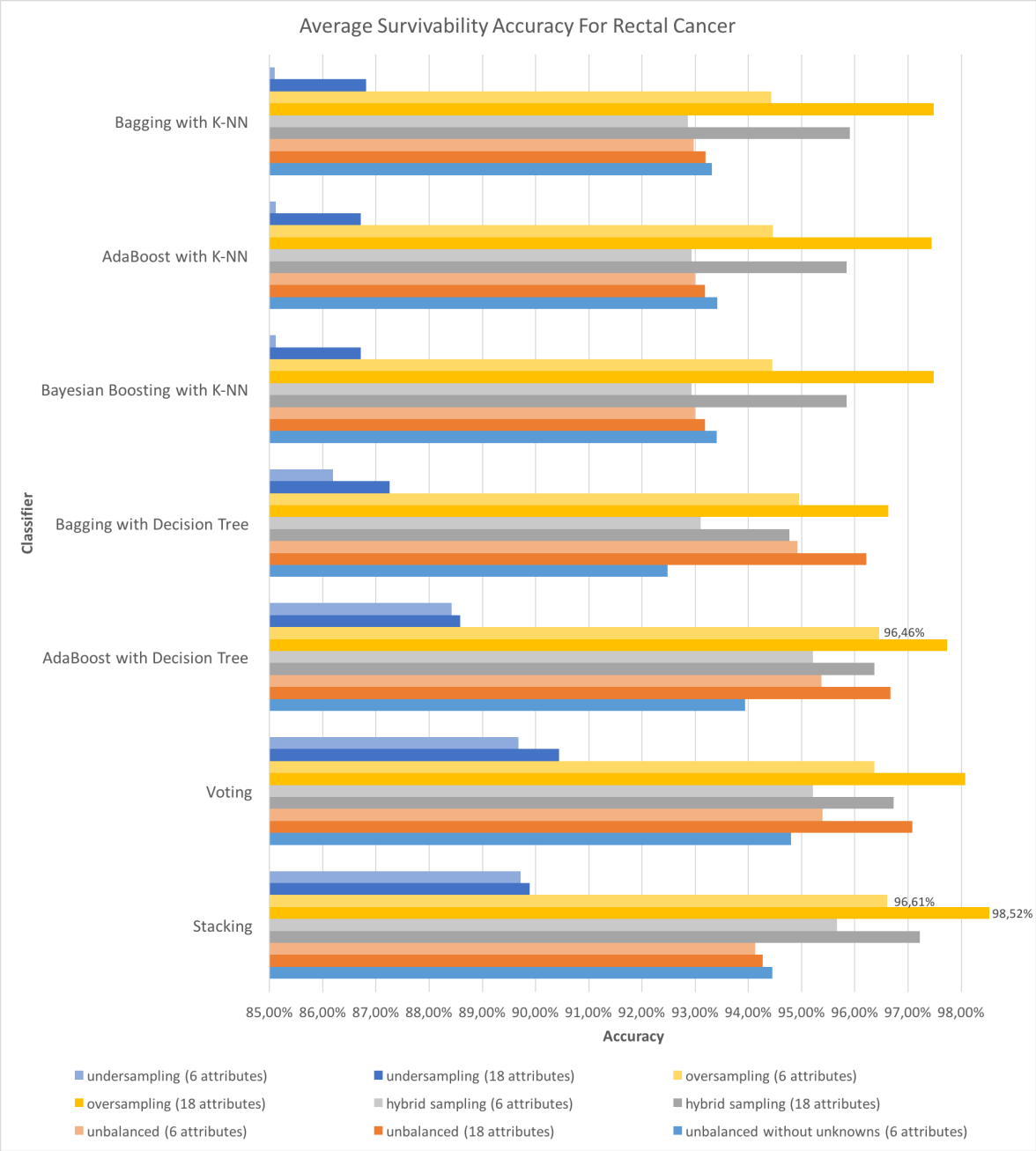



Figure 34.: Average survivability percentage accuracy for rectal cancer: comparison of the 18-attribute models with the 6-attribute models for the best learning schemes.

Figure 37 shows the average percentage of wrongly classified cases of the best learning schemes for the 5 years for the models trained with unbalanced datasets, unbalanced datasets without unknowns values, balanced oversampled datasets, balanced undersampled datasets and hybrid balanced datasets, for 6 and 18 attributes.

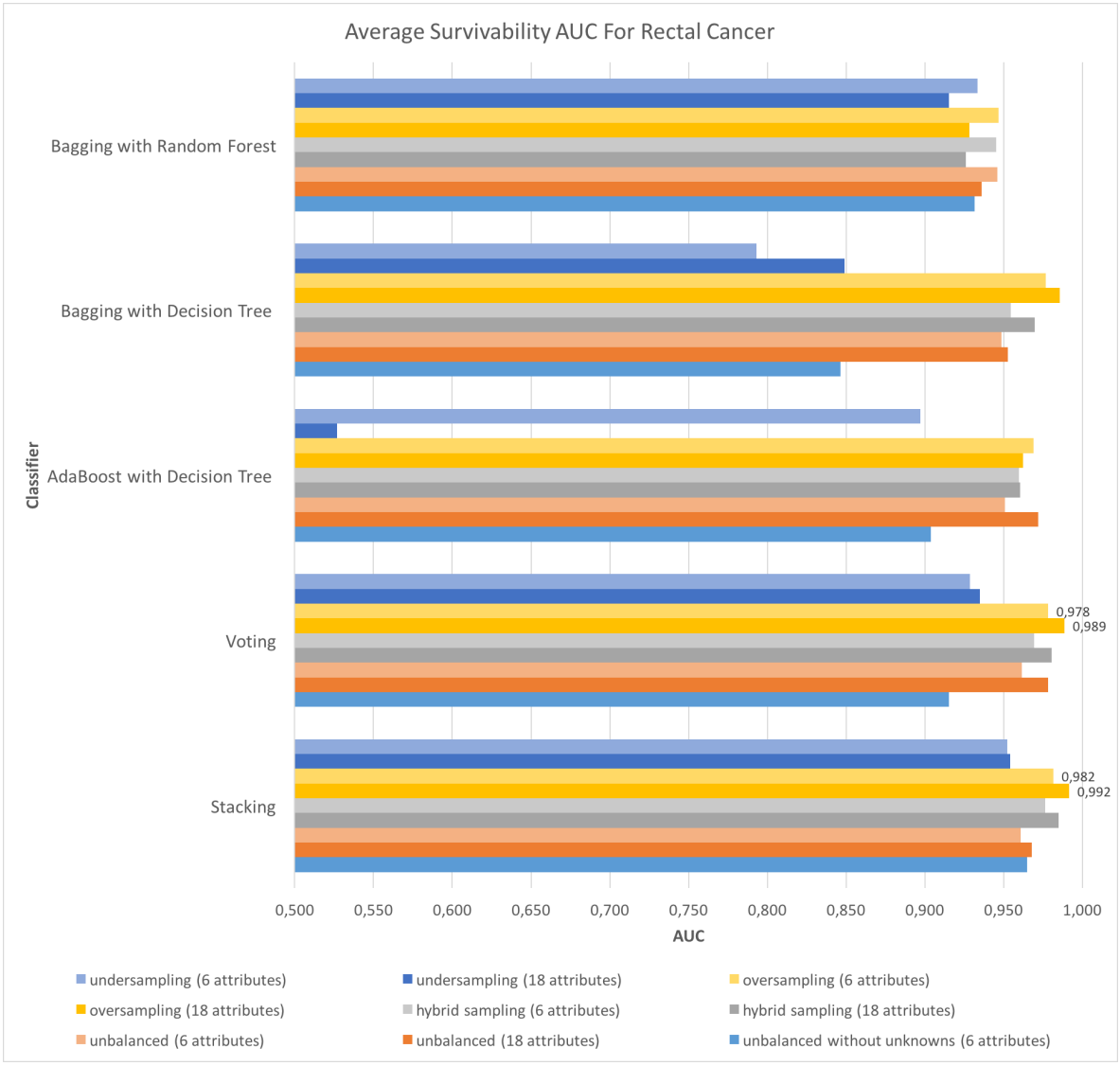



Figure 35.: Average survivability AUC for rectal cancer: comparison of the 18-attribute models with the 6-attribute models for the best learning schemes.

## 4.2 Conditional Survival Prediction Models

### 4.2.1 *Colon Cancer*

Table 19 shows the performance values for the conditional survival prediction models of colon cancer.

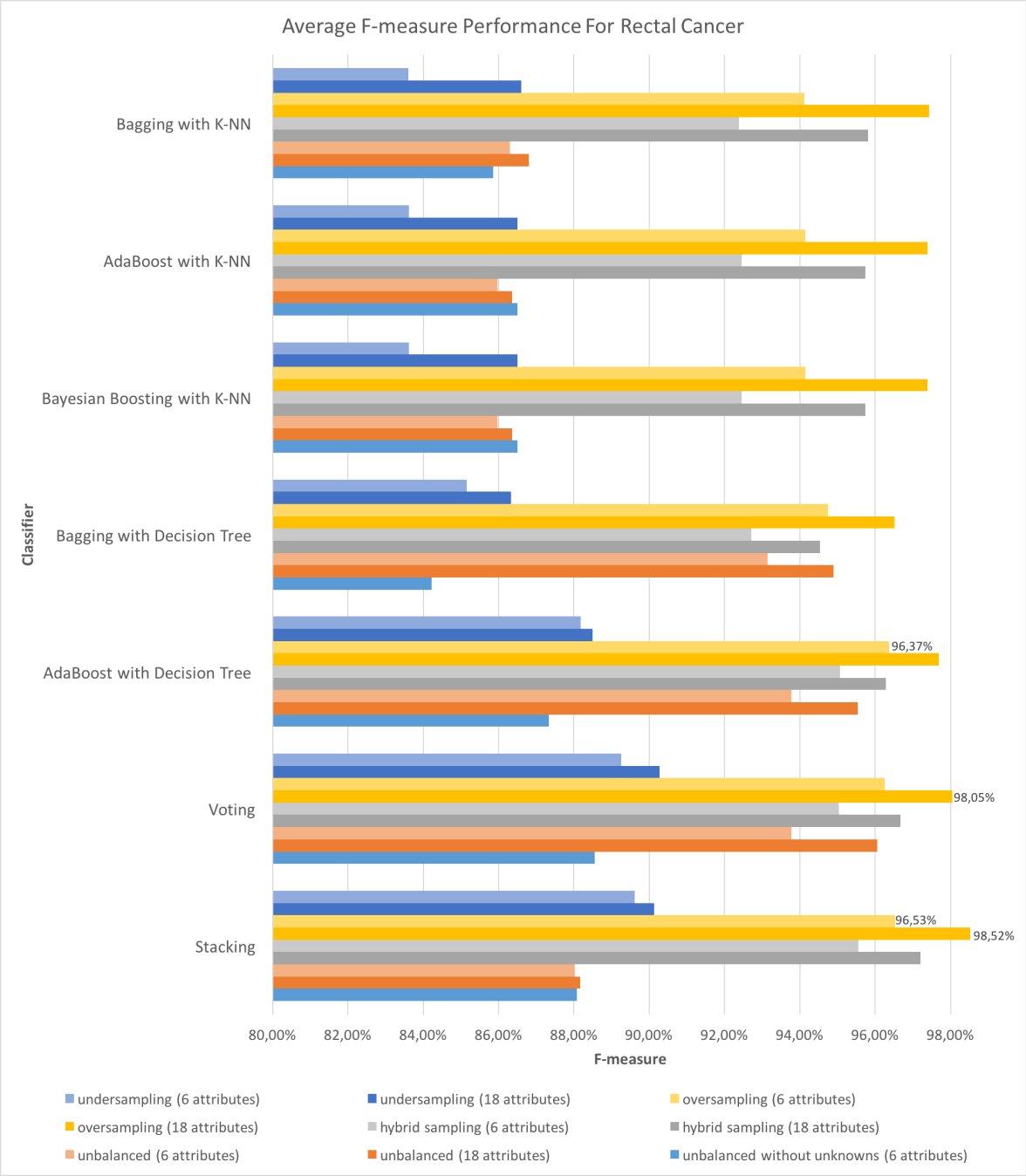



Figure 36.: Average F-measure performance for rectal cancer: comparison of the 18-attribute models with the 6-attribute models for the best learning schemes.

### 4.2.2 *Rectal Cancer*

Table 20 shows the performance values for the conditional survival prediction models of rectal cancer.

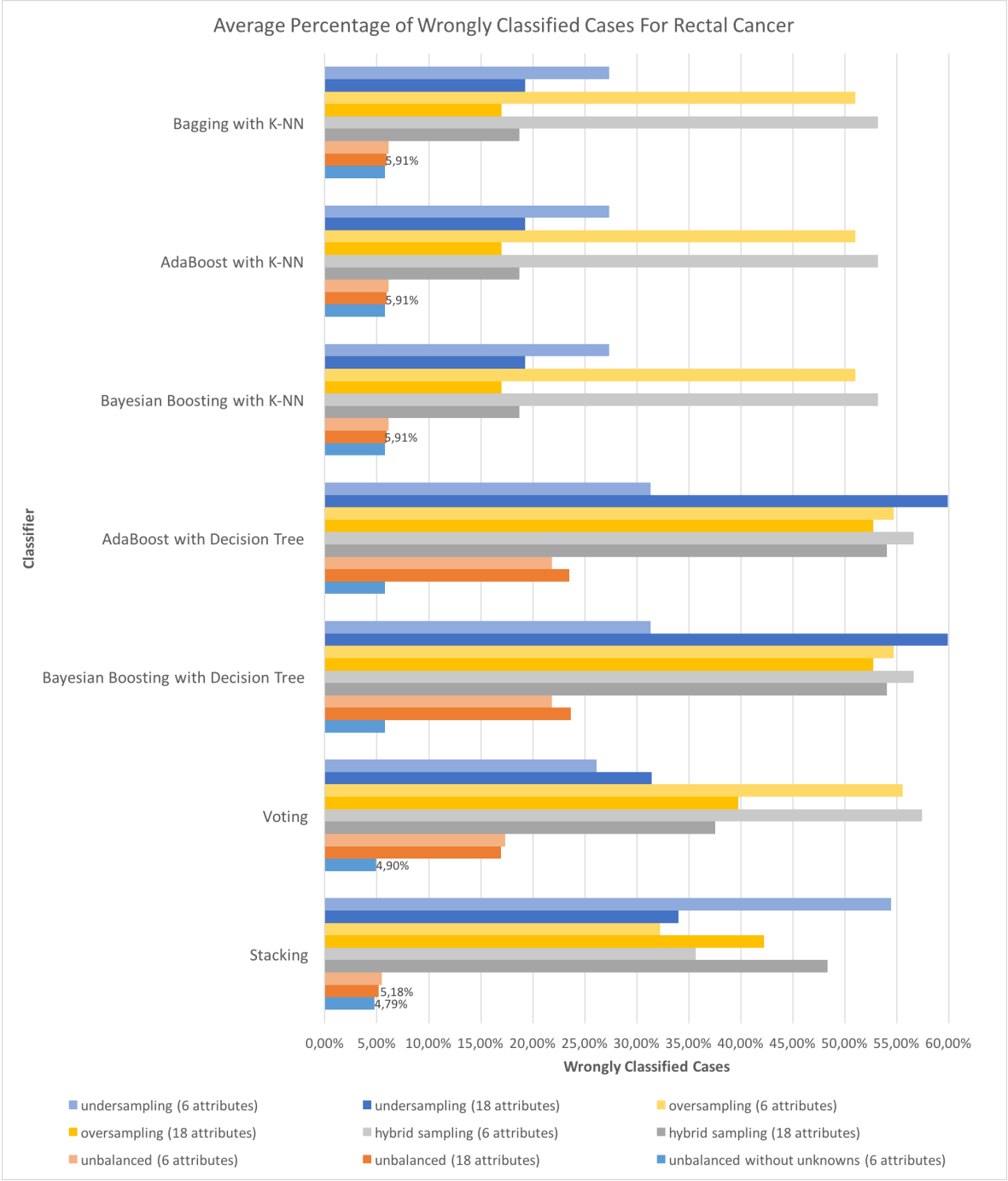



Figure 37.: Average percentage of wrongly classified cases for rectal cancer: comparison of the 18-attribute models with the 6-attribute models for the best learning schemes.

Table 19.: Performance values for the conditional survival prediction models of colon cancer.

| | Target Labels | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | survived the 1st year | | | | survived the 2nd year | | | survived the 3rd year | | survived the 4th year |
| | 2 Year | 3 Year | 4 Year | 5 Year | 3 Year | 4 Year | 5 Year | 4 Year | 5 Year | 5 Year |
| Accuracy | 97.39% | 97.27% | 97.44% | 97.02% | 98.32% | 97.62% | 97.09% | 98.48% | 97.55% | 98.59% |
| AUC | 0.981 | 0.985 | 0.986 | 0.984 | 0.979 | 0.983 | 0.979 | 0.974 | 0.969 | 0.945 |
| F-measure | 88.74% | 92.72% | 98.35% | 98.03% | 89.77% | 98.64% | 98.29% | 99.20% | 98.68% | 99.27% |
| Wrongly Classified Cases (%) | 2.57% | 3.08% | 2.45% | 3.08% | 1.80% | 2.58% | 3.61% | 1.62% | 2.60% | 1.10% |
| Total number of testing cases | 1752 | | | | 1552 | | | 1424 | | 1362 |

Table 20.: Performance values for the conditional survival prediction models of rectal cancer.

| | Target Labels | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | survived the 1st year | | | | survived the 2nd year | | | survived the 3rd year | | survived the 4th year |
| | 2 Year | 3 Year | 4 Year | 5 Year | 3 Year | 4 Year | 5 Year | 4 Year | 5 Year | 5 Year |
| Accuracy | 96.31% | 95.41% | 94.59% | 94.03% | 98.24% | 96.16% | 94.76% | 97.21% | 96.27% | 97.99% |
| AUC | 0.952 | 0.957 | 0.952 | 0.947 | 0.947 | 0.941 | 0.939 | 0.884 | 0.912 | 0.874 |
| F-measure | 70.13% | 78.99% | 96.87% | 96.45% | 79.15% | 97.91% | 97.09% | 98.56% | 98.02% | 98.97% |
| Wrongly Classified Cases (%) | 3.67% | 3.05% | 6.31% | 6.31% | 2.61% | 4.35% | 5.22% | 2.05% | 2.51% | 2.12% |
| Total number of testing cases | 491 | | | | 460 | | | 438 | | 425 |

## 4.3 DISCUSSION

The total of the 18 attributes were indicated by the expert physician as being shared for the survivability prediction for both cancers. The Feature Selection process picked 6 distinct attributes, for each type of cancer. The selected attributes, with the exception of the site-specific factors (the interpretation of the highest CEA test results and the clinical assessment of regional lymph nodes) for colon cancer and the surgical procedure for rectal cancer, they were all connected with the features indicated by the specialist physician. Still within the Feature Selection process, all the selected attributes had the same weight. It can be a limitation of the used software. Although when using ensemble learning to train models, each basic classifier in a meta-classifier is assigned a weight.

From a general observation of the results, for both types of cancer, the ensemble models using decision trees (excepting the Bayesian Boosting modeling) and k-NN demonstrated a very high performance values. Stand out all values of wrongly classified cases are the same, inside of each cancer type and dataset type, for models which only used k-NN classifier in the learning process.

For colon cancer, the model which presented the general highest values of performance (values from cross-validation) was the Stacking model trained with balanced oversampled dataset, with the average values of 98.23% and 97.04% for accuracy, 0.992 and 0.990 of AUC and 98.22% and 97.00% of F-measure, for 18 and 6 attributes, respectively. However, when the testing data was applied to this model and the prediction values were compared to the real values the percentage of wrongly classified cases was 19.38% and 56.78%, for 18 and

6 attributes, respectively. These values compared with the values of the models trained with unbalanced datasets are very high. The best model, in terms of wrongly classified cases percentage, was the Stacking model trained with the unbalanced dataset and without unknown cases (3.03% for 6 attributes). The values for 18 attributes of this type of dataset were not able to be calculated due the insufficient number of registries after remove the unknown values. Comparing the performance values of the Stacking model of the balanced oversampled dataset and the Stacking model of the unbalanced without unknown cases dataset, they are very close. For the 6 attributes, the model of the oversampled dataset was slightly better in 0.34% of accuracy, 0.003 of AUC and 1.84% of F-measure.

Still for the colon cancer, comparing the values of performance between models with 18 and 6 attributes, the 18 attributes models showed a slightly better performance values. However, it should be noted that, in addition to the close performances, the difference between the number of attributes used is important. To apply the attributes in a practical way (for instance, into a tool), the health care professional will lose much time if he must introduce 18 attributes. This is a critical point, may lead to non-use of the tool. The results show that it is possible to build a model with less than half of the features indicated by the expert physician, with similar performances.

For rectal cancer, the observed results were very identical. However, slightly lower. The model which presented the general highest values of performance (values from cross-validation) also was the Stacking model trained with balanced oversampled dataset, with the average values of 98.52% and 96.61% for accuracy, 0.992 and 0.982 of AUC and 98.52% and 96.53% of F-measure, for 18 and 6 attributes, respectively. The values of wrongly classified cases percentage were 42.25% and 32.22%. The best model, in terms of wrongly classified cases percentage, was the Stacking model trained with the unbalanced dataset and without unknown cases (4.79% for 6 attributes). The difference of performance values of the Stacking model trained with the balanced oversampled dataset and the Stacking model trained with the unbalanced dataset and without unknown cases, were slightly higher relative to the same obtained values for colon cancer. The Stacking model trained with the balanced oversampled dataset was better in 2.16% of accuracy, 0.017 of AUC and 8.45% of F-measure.

As mentioned throughout the dissertation, the intention was to develop the conditional survival models using the same conditions which the best found model for the prediction survival after treatment models. Taking into account the performance values (from cross-validation), the percentage of wrongly classified cases and their relations with the number of attributes, the Stacking model trained with the unbalanced dataset and without unknown cases was selected to develop the conditional survival models for colon and rectal cancer, using the 6 selected attributes by Feature Selection process. These models presented high

values of performance and a low percentage of wrongly classified cases, similar to values obtained for the prediction survival after treatment models, for each type of cancer.

Comparing this approach with others mentioned in Chapter 2, for colon cancer, fewer features were necessary to develop the prediction model. Moreover, in the approach followed in [3], the closest to the one followed herein for the prediction after treatment, the best model of colon cancer survival prediction was based on a Voting classification scheme, with prediction accuracies of 90.38%, 88.01%, and 85.13% and AUCs of 0.96, 0.95, and 0.92 for years 1, 2 and 5. In this work, the model considered better presented less attributes and a larger time window. The performance values for each year also were always better. With 95.85%, 96.74%, 96.88%, 97.01% and 97.03% of accuracy and 0.982, 0.985, 0.988, 0.988 and 0.988 of AUC for years 1 to 5.

To calculate the conditional survival for colon cancer patients, the best model in Chapter 2, presented by Chang et al. [18] had a C-index of 0.816 (a model from 0 to 5 years). The average of AUC for the correspondent models in this work was 0.977, presenting one characteristic more.

Relative to rectal cancer, the best model in Chapter 2 for prevision after treatment had a C-index of 0.70 (from 1 to 10 years) by Valentini et al. [98]. In this work, the model considered better presented three less attributes and the AUC values were 0.957, 0.968, 0.968, 0.968 and 0.963, for 1 to 5 years, respectively. Producing a average of 0.965, also a higher value than the existing solution.

The last but not the least, to calculate the conditional survival of rectal cancer patients, the actual solution presents a value of 0.75 of C-index (from 0 to 5 years). The corresponding developed model, in this work, produced a average of 0.931 of AUC.

As such, the present work represents an improvement and was able to achieve considerably better results.

# 5

# DEVELOPMENT OF AN APPLICATION

Over the last years, the industry of wireless devices, such as handheld PCs and even smartphones, gone through tremendous advancements. Accordingly, these devices have become increasingly popular and common [88]. Throughout the last decade, mobile phones have gone from being simple phones to being handheld pocket-sized computers. Their powerful capabilities, since the processing and on-board computing capacity till the high quality of screens, incite the development of applications [8].

Statistics shows that in 2016 the number of smartphone users is forecast to reach 2,08 billion. It is projected that just over 36% of the world population will use a smartphone by 2018, up from about 10% in 2011 [67]. According to data from the International Data Corporation (IDC) Worldwide Quarterly Mobile Phone Tracker, the Android of Google and iOS of Apple are the two most popular smartphone operating systems [22].

For the health care industry, mobile applications yielded new boundaries in providing better care and services to patients. Moreover, it is making a revolution in the way of the information is managed in this industry and redefine the doctor – patient communication [88, 64]. Mobile devices and their applications have provided many benefits for health care professionals. The portability of mobile applications can increase the productivity of these professionals. It grants a rapid access to information and multimedia resources, allowing them to make decisions more quickly with a lower error rate, increasing the quality of patient documentation and improved workflow patterns [100, 59].

This chapter describes how the models were made available to health care professionals.

## 5.1 REQUIREMENTS GATHERING

The current section is intended to specify the functional and non functional requirements to construct an application able to help the health care professionals in their functions, specifically to help physicians to predict the survivability for colorectal patients. The specification of the requirements are represented in a textual mode. To the requirements gathering were used insight techniques and was analyzed of the dissertation proposal. However, it will not specify the technique used in each requirement.

### 5.1.1 *Functional Requirements*

The functional requirements describe the functionality that is expected the system to have. The application must:

- Allow the selection of the type of prediction;
- Allow the selection of the cancer type for which the user want to obtain a prediction;
- Allow the insertion of values for a set of selected characteristics to the prediction models;
- Allow the choice of the value to insert, for a characteristic, from a set of listed values;
- Allow the obtain of the survival prediction, according to the inserted data, for 1, 2, 3, 4 and 5 years after the diagnosis and treatment;
- Allow the obtain of the conditional survival prediction, according to the inserted data, for 2, 3, 4 or 5 years after the diagnosis and treatment, according to the years that patient already survived;
- Allow the obtain of a confidence value for each year of prediction;
- Allow the visualization of the prediction data through a bar chart;
- Allow the insertion of the selected characteristics of the patients into a new registry of a database;
- Not allow user submit the form of characteristics if some them have not any value selected;
- Not allow user submit the selected characteristics if the mobile device is not connected to the internet.

### 5.1.2 *Non-functional requirements*

To ensure a quality mobile tool, were defined some essential points that describe how the system works. These point are commonly known as non-functional requirements and are related with the use of the application in terms of performance, usability, reliability, security, availability, maintenance and technologies involved. The non-functional requirements recognized are:

- Available of a mobile application;

- Available of the mobile application using open source technologies, constantly update and evolution;
- The mobile application should have responsive design and to adapt to different mobile devices;
- The mobile application should have an appealing appearance;
- The visualization of the characteristics and predictions should be of a easy comprehension;
- The insertion of the values for the characteristics should be easy;
- The solution should cover the principal mobile platforms (iOS, Android and Windows Phone).

## 5.2 ARCHITECTURE

To made available the developed models was projected a hybrid mobile application, appropriate to smartphones and tablets. On the back-end of this tool, were developed two web services: one for give the survivability prediction response of colon or rectal cancer to user (applying the models) and another to recalculate the models. Figure 38 shows the architecture of the application. It will be discussed in more detail in the next subsections.

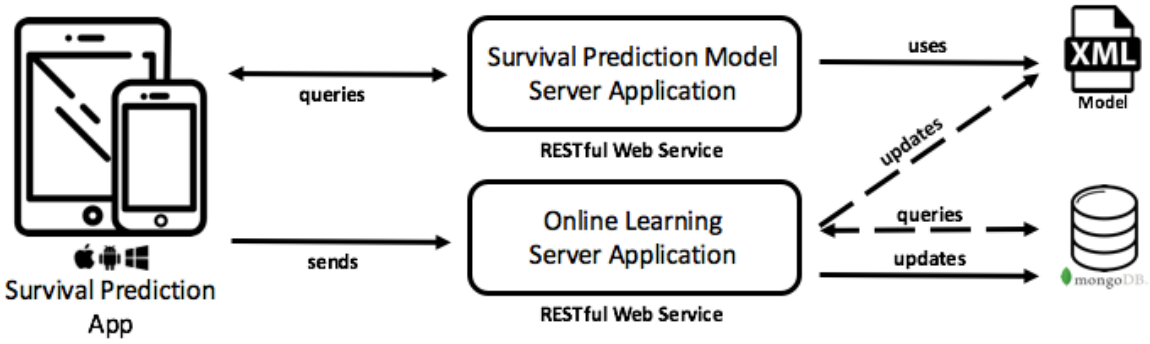



Figure 38.: Architecture of the developed tool.

### 5.2.1 *Survival Prediction Application*

Hybrid approach fits between web and native methodology. To construct a native application is necessary know the correspondent programing language, which vary according to the operating system target. In the case of iOS it is Objective-C or Swift, for Android it

is Java and for Windows Phone it is C#. In a web-based approach, the mobile device will not have any application specific components installed. The applications are browser based and are platform independent. However, the accessability could be conditioned because it is exposed to cross space communication vulnerabilities [71]. For that reason, it cannot be appropriate to frequent uses. A hybrid application is developed applying web technologies (mainly, HTML5, CSS and JavaScript) and gets executed inside native container on the mobile device. It is suitable to multiple platforms and is distributable through an application store, like native applications. This type of approach can have an inferior performance, comparing with native applications [71]. However, mobile devices in nowadays have powerful capabilities and the lapse of performance are not really noted.

The application was created with intention to be a cross platform tool, suitable to smartphones and tablets, either with the android, Windows Phone or even with the iOS operating system.

The application was developed using AngularJS, Ionic Framework, and Cordova. AngularJS is a popular framework of JavaScript, mainly maintained by Google [5]. Ionic is an HTML5 SDK open source, which offers a library of mobile-optimized HTML, CSS and JavaScript CSS components, gestures, and tools for building interactive apps [40]. It also was optimized for AngularJS. Cordova wraps the HTML/JavaScript app into a native container which can access the device functions of several platforms [21]. These functions are exposed via a unified JavaScript API, for an easily accessing to the full native functionality.

### 5.2.2 *Survival Prediction Model Server Application*

In order to apply the developed models, was developed a web service with representational state transfer (REST) architecture. The RESTful architecture style is based on web-standards and the HTTP protocol. It was chosen due to the declarative nature and other characteristics of it, such as being light-weight, easily accessible and scalable [107].

The web service was developed in Java with the Java API for RESTful Web Services (JAX-RS) [45] reference implementation Jersey [41], an open source framework used for Java projects and distributed mainly via Maven [55].

The data is sent over the HTML POST method when the health care professional submits the values of characteristics for a particular patient through the application. The RESTful web service, integrating the RapidMiner software through its API, receives the values and applies them to the corresponding model, which is in a XML file. The response of the patient survivability for 1 to 5 years is returned in a JSON format and a bar chart is constructed.

### 5.2.3 *Online Learning Server Application*

The submitted data from each patient and its outcomes are added to a database for posteriorly, after several insertions, recalculate all the models, keeping them up-to-date. Thereunto, was developed another RESTful web service, with a similar structure to that developed for patient survivability respond. It provides an independent service because of the time necessary for recompute the models.

The data is inserted into a NoSQL database (MongoDB [56]) and for each 10% new registries (in relation of the initial cases) the models for 1 to 5 years are recalculated, generating 5 XML files which replace the used files in 5.2.2. In this way, it will be possible for users to dynamically feed new cases to the prediction system and make it change in order to provide better survival predictions. This type of model could also prove to be very useful when integrated in computer-interpretable guideline systems, such as the one described in [61], as a way to provide dynamic knowledge to rule-based decision support.

## 5.3 INTERFACE

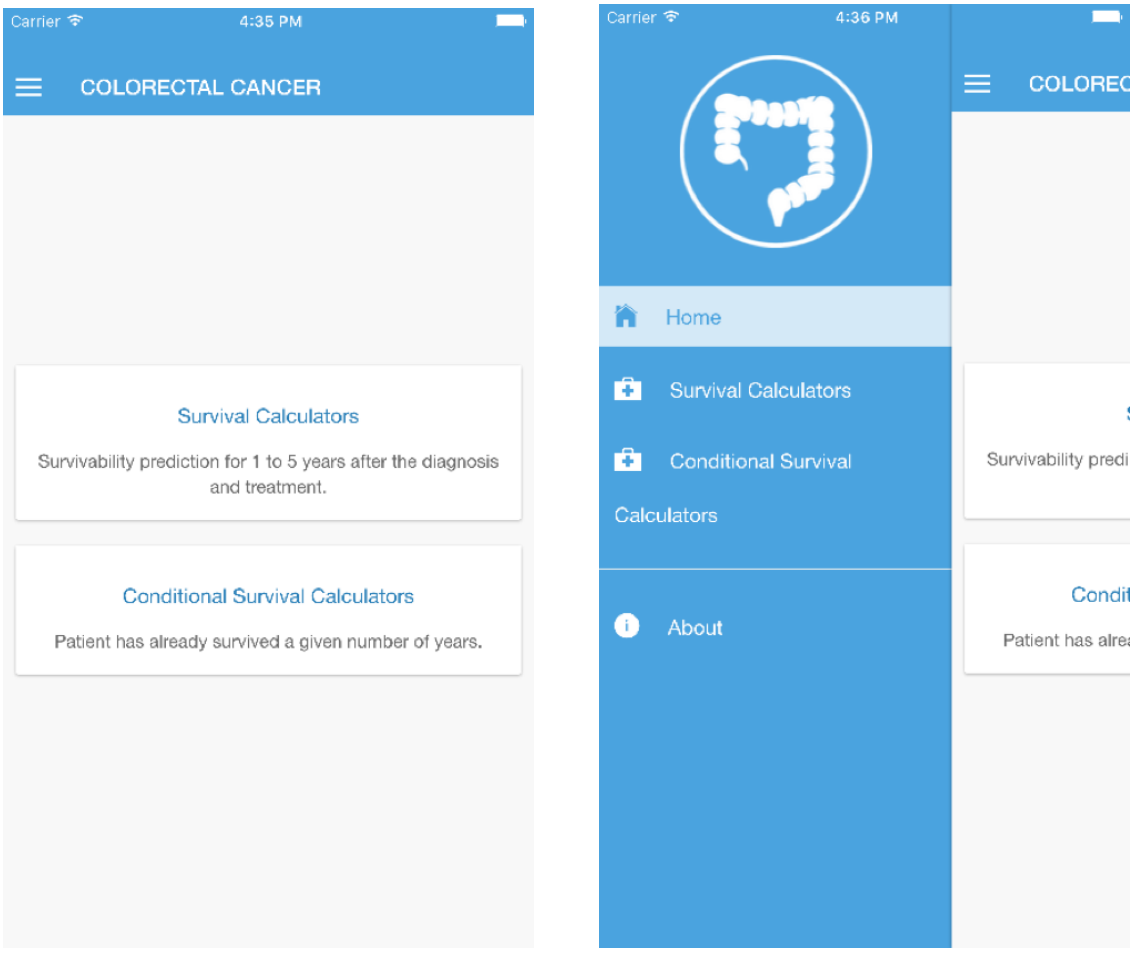
Figure 39a shows the first screen which appears when the application is initiated. Here, user has the opportunity to choose the type of prediction he wants do: calculate the survivability after treatment or calculate the conditional survival. Clicking on the menu (Figure 39b) are visible all the options available in this application.

Choosing calculate the survivability after treatment (Figure 40a) or calculate the conditional survival (Figure 40b), a new menu comes. Here, user has the occasion to select the cancer type for the prediction.

## 5.4 USE CASE

### 5.4.1 *Survivability After Treatment Calculators*

A typical use case is getting a prediction for colon cancer survivability. Supposing a physician is treating a patient diagnosed with colon cancer, once the type of cancer in the home screen is set (as shown in Figure 39b), the health care professional inserts the values for the selected features (Figure 41a). All features, except for the age of the patient, are filled in by choosing the value from a list of available options. By submitting a case of a patient with *55* years old, having a *positive/elevated* carcinoembryonic antigen value, with clinical assessment of regional lymph nodes of *not clinically evident*, with the primary site of the cancer being in the *sigmoid colon*, with stage *0* and with *5* as the number of regional nodes examined, the values are sent to the *Survival Prediction Model Server Application* and the

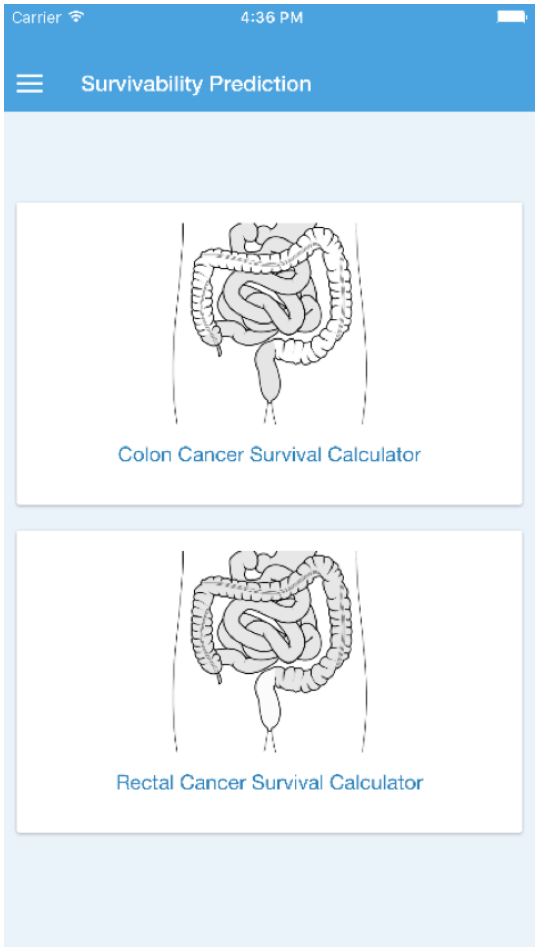



(a) Home screen. (b) Menu.

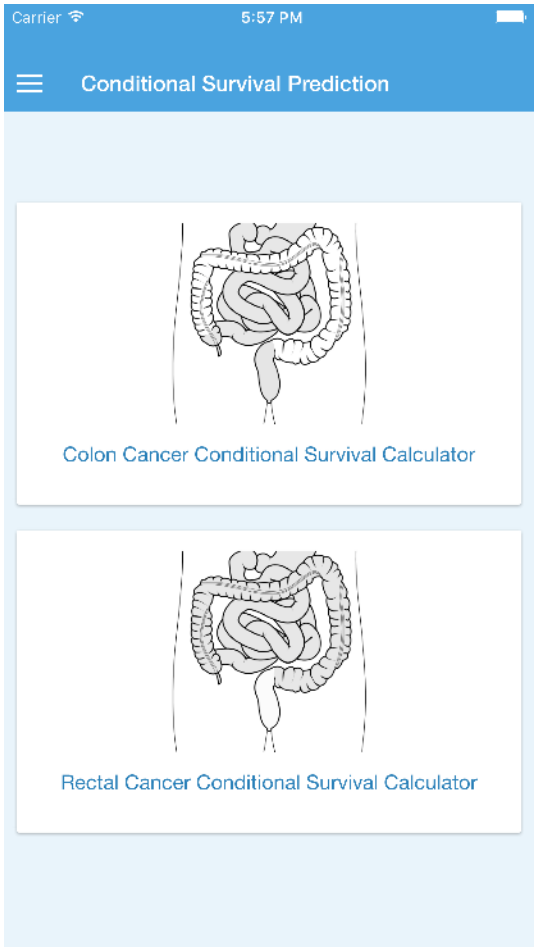
Figure 39.: Home screen and menu of the application.

outcome is calculated. The prediction is always provided in the form of confidence values for a positive prediction, i.e., the confidence that the patient will survive. This is displayed in a new screen in the form of a bar chart (Figure 41b). For the stage of the patient, the physician can choose between the TNM system or the grouped stage, known as AJCC stage. The results show that, while the model was able to predict with 100% confidence that the patient will survive the first three years, the confidence of his surviving the fourth and fifth years is 0%.

To predict the survivability of a patient diagnosed with rectal cancer, the procedure is similar to the one used for colon cancer. Figure 42a shows what happen when the user not filled out all the characteristics and Figure 42b shows what happen when the internet connection fails, i.e., the submit button of the application is not enabled.




(a) Survivability after treatment calculator.



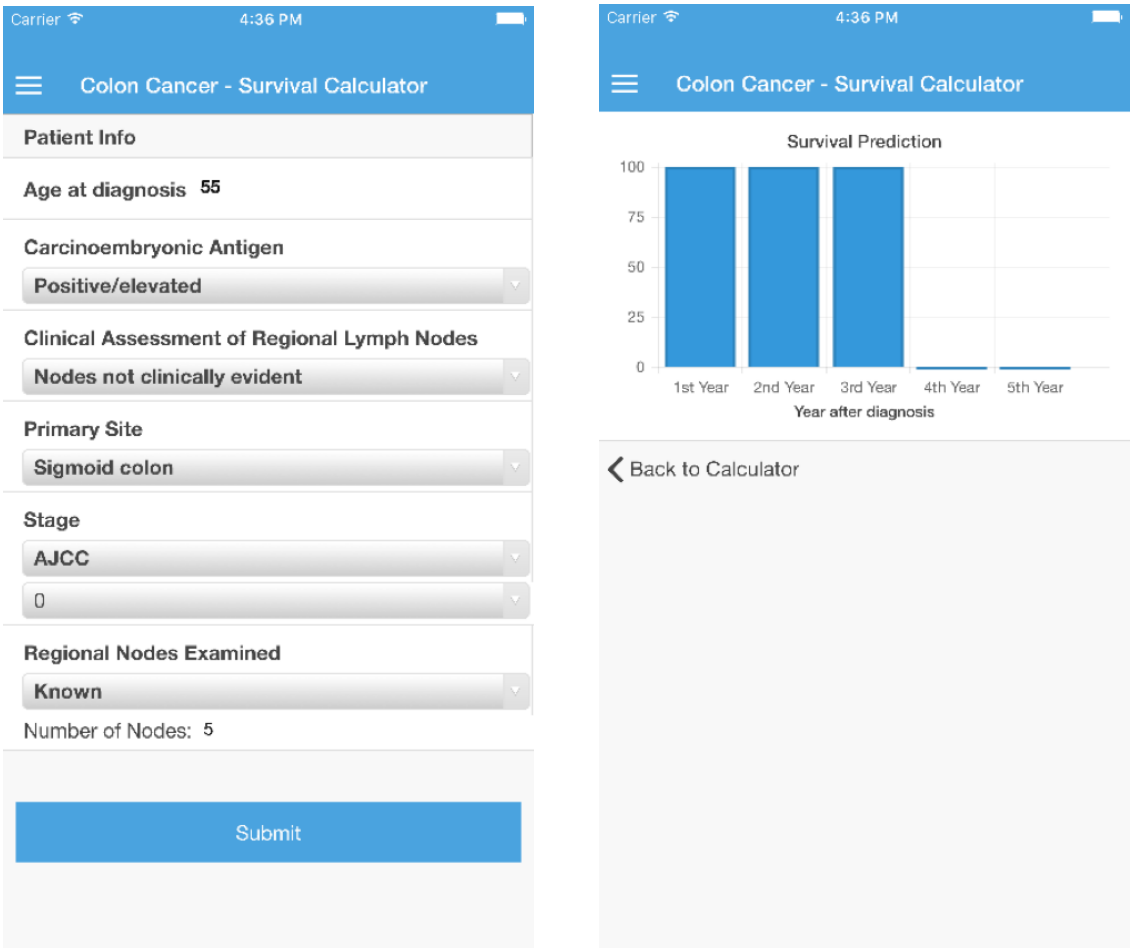

(b) Conditional survival calculator.

Figure 40.: Calculator menus.

### 5.4.2 *Conditional Survival Calculators*

To predict the survivability of a patient who already survived some time after the diagnosis and treatment for colon or rectal cancer, the procedure is similar to the one used in survival after treatment calculators. The only difference is the selection of the year that patient already survived. Supposing a physician wants to know the survival of the patient diagnosed with colon cancer, after he/she survived the first year after treatment. The health care professional inserts the values for the features (Figure 43a). The results (Figure 43b) show that, the model was able to predict, one year after treatment of colon cancer, with 100% confidence that the patient will survive more two years. The confidence for the fourth and fifth years after treatment, remained at 0%.

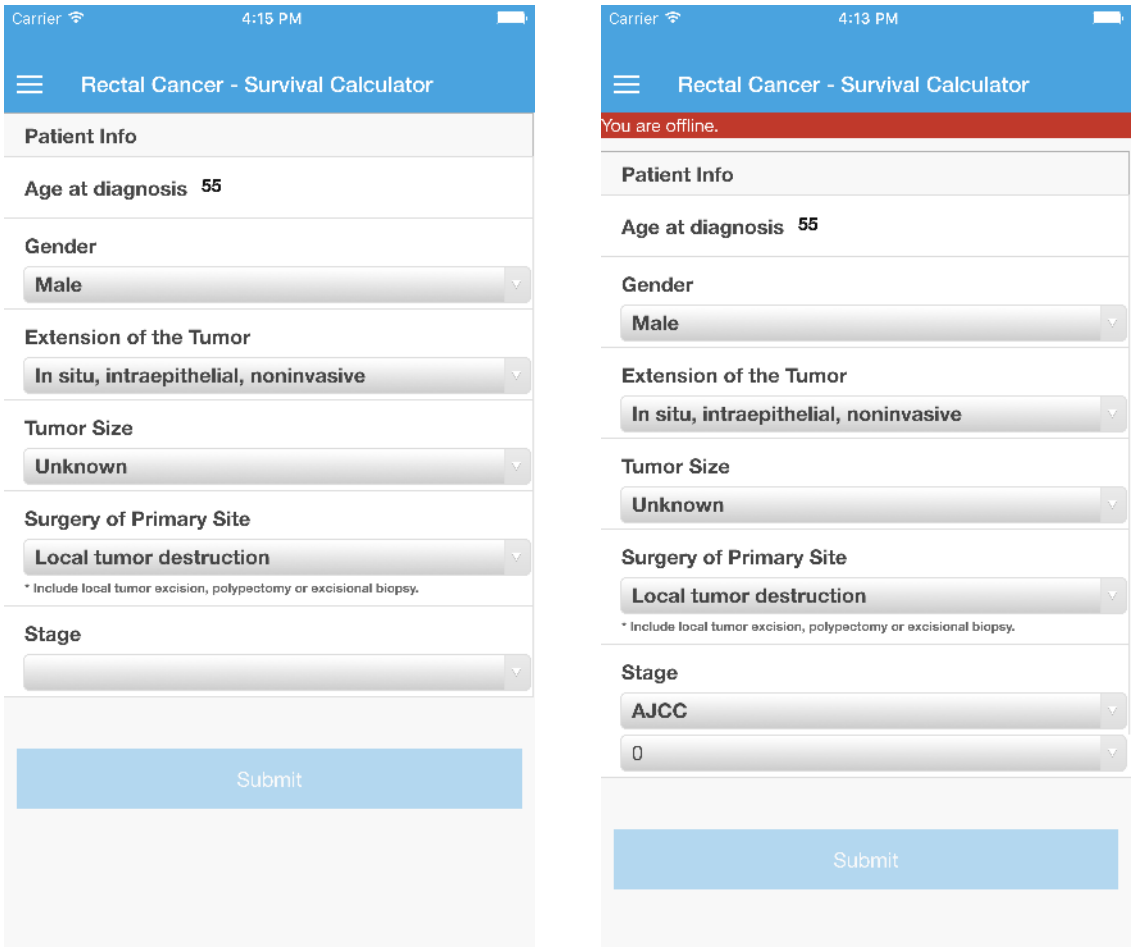
To predict the conditional survival of a patient diagnosed with rectal cancer, the procedure is similar to the one used for the conditional survival of colon cancer patients.




(a) Characteristics of Colon Cancer.

(b) Results for Colon Cancer.

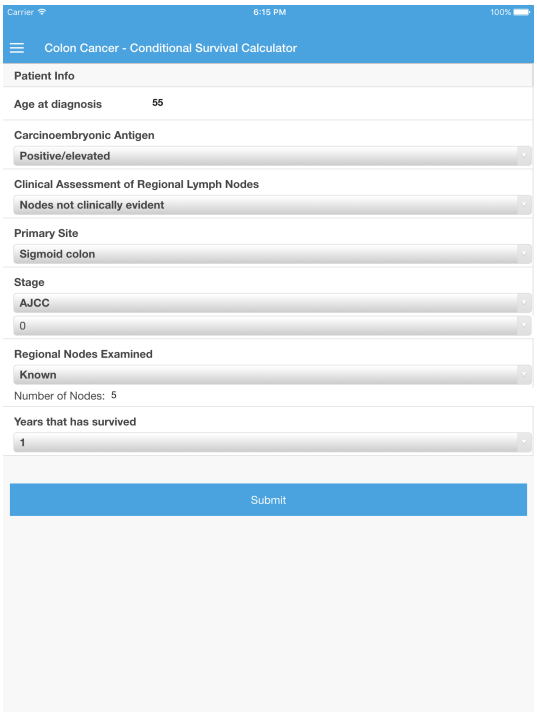
Figure 41.: Colon Cancer Survivability After Treatment Calculator (accessed by a smartphone).



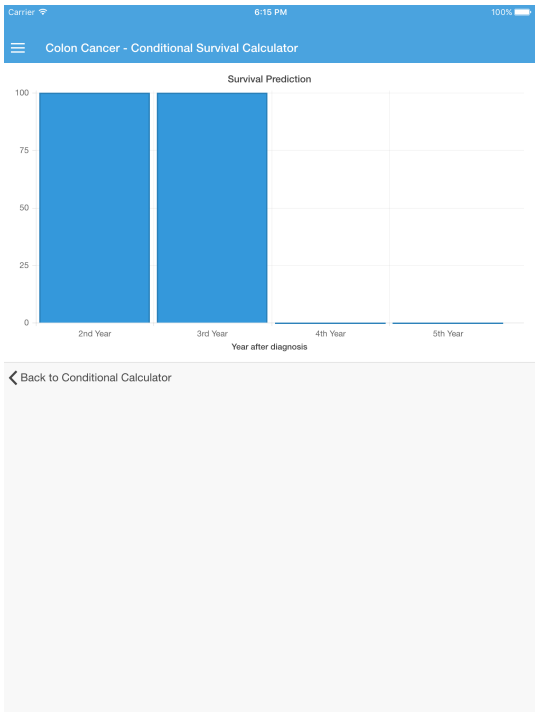

(a) Characteristics are not filled.

(b) Internet connection fail.

Figure 42.: Error control of application.




(a) Characteristics of Colon Cancer (Conditional Survival).




(b) Results for Colon Cancer (Conditional Survival)).

Figure 43.: Colon Cancer Conditional Survival Calculator (accessed by a tablet).

# 6

# CONCLUSIONS, PUBLICATIONS AND FUTURE WORK

## 6.1 CONCLUSIONS

This work involved the use of different meta-classification schemes to construct survival prediction models, for colon and rectal cancer patients after treatment and after some time after treatment. The best models found was a Stacking classification scheme, combining k-NN, Decision Tree, and Random Forest classifiers as base learners and a Naive Bayes classifier as a stacking model learner. These models were trained with an unbalanced dataset, without unknown cases for the selected features. Therefore, the objectives 1, 2 and 3 were fulfilled with success.

The ideal number of features for colon and rectal cancer survival prediction was found to be 6. The selected set for colon cancer includes: age, CS site-specific factor 1, CS site-specific factor 2, derived AJCC stage group, primary site, and regional nodes examined. For rectal cancer, the selected set was: age, CS extension, CS tumor size, derived AJCC stage group, RX summ–surg prim site and the sex of patient at time of diagnosis. The selected attributes for colon and rectal cancer were not all the same, only the age and the stage were common. It was observed that the most of the selected features were connected with the features indicated by the specialist physician. It was also observed that the selected features are less than half of the features given by specialist physicians (18 attributes), and presented similar performance values. In this manner, the objectives 4, 5 and 6 were fulfilled with success.

Many studies [19, 50] show how important the problem of using imbalanced datasets is, from both the algorithmic and performance perspectives. During this work, three sampling forms were tested and the results were compared with unbalanced datasets. Was concluded that the sampling improved the performance values. However, the models trained with balanced dataset were those that classified worst. Concluding, in this manner, that not always the best models (with better performance values) are those that classify better.

The developed models were able to present a better performance than the existing approaches and overall with fewer features. It was determined that not all models are avail-

able to health professionals and those that are, only two has the minimum characteristics to be considered mobile friendly. Consequently, the objective 7 was fulfilled with success.

The best developed models were available to health care professionals into a cross-platform mobile application, in order to assist them in carrying out their duties at any time. To ensure that the model is able to adapt and adjust, an online learning server was created. In this way, it will be possible for users to dynamically feed new cases to the prediction system and make it change in order to provide better survival predictions. Therefore, the objectives 8 and 9 were fulfilled with success.

## 6.2 publications

The current work originated two conference papers with the references:

- Ana Silva, Tiago Oliveira, Paulo Novais, José Neves and Pedro Leão. Developing an Individualized Survival Prediction Model for Colon Cancer. *7th International Conference on Ambient Intelligence*, 2016.
- Ana Silva, Tiago Oliveira, Vicente Julian, José Neves and Paulo Novais. A Mobile and Evolving Tool to Predict Colorectal Cancer Survivability. *12th IFIP International Conference on Artificial Intelligence Applications and Innovations*, 2016.

And a journal article with the reference:

- Ana Silva, Tiago Oliveira, Paulo Novais, José Neves and Pedro Leão. Treating Colon Cancer Survivability Prediction as a Classification Problem. *Advances in Distributed Computing and Artificial Intelligence Journal*, 5(1).

## 6.3 prospect for future work

Future work includes the continuous improving of the models and of the tool. The increase of the time windows, for the conditional survival calculators, is one of the proposals. In order to better understand comprehend the capacities of the developed tool, it will be interesting carrying out prospective tests.

Additionally, the adaptation of the models to Portuguese reality would be an important goal, because there is no Portuguese tool for this purpose. Compare a Portuguese tool with the current work would be enriching and would give possibility for physicians to prove, or not, that the lifestyle among people with different cultures influence the prediction for these types of cancer.

# REFERENCES


[1] Colon/rectum cancer — american cancer society. http://www.cancer.org/cancer/colonandrectumcancer/. (Accessed on 03/20/2016).

[2] F. E. Ahmed. Artificial neural networks for diagnosis and survival prediction in colon cancer. *Molecular Cancer*, 4(1):29, 2005.

[3] R. Al-Bahrani, A. Agrawal, and A. Choudhary. Colon cancer survival prediction using ensemble data mining on SEER data. pages 9–16, 2013.

[4] American Cancer Society. Cancer Facts & Figures 2015. Technical report, 2015.

[5] Angularjs — superheroic javascript mvw framework. https://angularjs.org. (Accessed on 31/03/2016).

[6] J. Betts. *Anatomy & physiology*. OpenStax College, Rice University, Houston, Texas, 2013.

[7] K. Bonewit-West, S. Hunt, and E. Applegate. *Today's Medical Assistant: Clinical & Administrative Procedures*. Elsevier Health Sciences, 2015.

[8] M. N. Boulos, S. Wheeler, C. Tavares, and R. Jones. How smartphones are changing the face of mobile and participatory healthcare: an overview, with example from ecaalyx. *Biomedical engineering online*, 10(1):24, 2011.

[9] T. L. Bowles, C.-Y. Hu, N. Y. You, J. M. Skibber, M. A. Rodriguez-Bigas, and G. J. Chang. Rectal cancer survival calculator. http://www3.mdanderson.org/app/medcalc/index.cfm?pagename=rectumcancer. (Accessed on 26/03/2016).

[10] T. L. Bowles, C.-Y. Hu, N. Y. You, J. M. Skibber, M. A. Rodriguez-Bigas, and G. J. Chang. An individualized conditional survival calculator for patients with rectal cancer. *Diseases of the colon and rectum*, 56(5):551–9, 2013.

[11] A. P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.

[12] L. Breiman. Bagging Predictors. *Machine Learning*, 24:123–140, 1996.

[13] D. M. Bush and J. S. Michaelson. Colon cancer outcome calculator. http://www.lifemath.net/cancer/coloncancer/outcome/index.php. (Accessed on 26/03/2016).

[14] D. M. Bush and J. S. Michaelson. Derivation : Nodes + PrognosticFactors Equation for Colon Cancer accuracy of the Nodes + PrognosticFactors equation . Technical report, 2009.

[15] Cancer Research UK. http://www.cancerresearchuk.org, 2016. (Accessed on 22/02/2016).

[16] D. Carneiro, R. Costa, P. Novais, J. Neves, J. Machado, and J. Neves. Simulating and Monitoring Ambient Assisted Living. In *Proceedings of the ESM 2008 - The 22nd annual European Simulation and Modelling Conference*, pages 175–182, Le Havre, 2008.

[17] G. J. Chang, C. Y. Hu, C. Eng, J. M. Skibber, and M. A. Rodriguez-Bigas. Colon cancer survival calculator. http://www3.mdanderson.org/app/medcalc/index.cfm?pagename=coloncancer. (Accessed on 26/03/2016).

[18] G. J. Chang, C. Y. Hu, C. Eng, J. M. Skibber, and M. A. Rodriguez-Bigas. Practical application of a calculator for conditional survival in colon cancer. *Journal of Clinical Oncology*, 27(35):5938–5943, 2009.

[19] N. V. Chawla. Data Mining for Imbalanced Datasets: An Overview. In *Data Mining and Knowledge Discovery Handbook*, pages 853–867. 2005.

[20] E. Coiera. *Guide to medical informatics, the internet, and telemedicine*. Chapman & Hall Medical, London New York, 1997.

[21] Apache cordova. http://cordova.apache.org. (Accessed on 31/03/2016).

[22] I. D. Corporation. Idc: Smartphone os market share 2015, 2014, 2013, and 2012. http://www.idc.com/prodserv/smartphone-os-market-share.jsp. (Accessed on 30/03/2016).

[23] Â. Costa, P. Novais, J. M. Corchado, and J. Neves. Increased performance and better patient attendance in an hospital with the use of smart agendas. *Logic Journal of IGPL*, 2011.

[24] R. Costa, P. Novais, J. Neves, G. Marreiros, C. Ramos, and J. Neves. *VirtualECare: Group Decision Supported by idea Generation and Argumentation*, pages 293–300. Springer US, Boston, MA, 2008.

[25] S. Džeroski and B. Ženko. Is combining classifiers with stacking better than selecting the best one? *Machine Learning*, 54(3):255–273, 2004.

[26] P. F. Engstrom, J. P. Arnoletti, A. B. Benson, Y.-J. Chen, M. a. Choti, H. S. Cooper, A. Covey, R. a. Dilawari, D. S. Early, P. C. Enzinger, M. G. Fakih, J. Fleshman, C. Fuchs,

J. L. Grem, K. Kiel, J. a. Knol, L. a. Leong, E. Lin, M. F. Mulcahy, S. Rao, D. P. Ryan, L. Saltz, D. Shibata, J. M. Skibber, C. Sofocleous, J. Thomas, A. P. Venook, and C. Willett. NCCN Clinical Practice Guidelines in Oncology: colon cancer. *Journal of the National Comprehensive Cancer Network : JNCCN*, 7:778–831, 2009.

[27] T. Fawcett. Roc graphs: Notes and practical considerations for researchers. *Machine learning*, 31(1):1–38, 2004.

[28] T. Fawcett. An introduction to {ROC} analysis. *Pattern Recognition Letters*, 27(8):861 – 874, 2006. {ROC} Analysis in Pattern Recognition.

[29] J. Ferlay, I. Soerjomataram, M. Ervik, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. Parkin, D. Forman, F. Bray, and F. Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide. Technical report, 2013.

[30] Y. Freund and L. Mason. The alternating decision tree learning algorithm. *International Conference on Machine Learning*, 99:124–133, 1999.

[31] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997.

[32] Y. Freund and R. R. E. Schapire. Experiments with a New Boosting Algorithm. *International Conference on Machine Learning*, pages 148–156, 1996.

[33] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting, 2000.

[34] M. Goel, P. Khanna, and J. Kishore. Understanding survival analysis: Kaplan-meier estimate. *International journal of Ayurveda research*, 1(4):274, 2010.

[35] J. HAINCE, G. Beaudry, G. GARON, M. Houde, T. Holzer, M. Beaulieu, and N. Bertrand. Method for detecting metastasis of gi cancer, 2012. EP Patent App. EP20,100,745,780.

[36] M. a. Hall. Correlation-based Feature Selection for Machine Learning. *Methodology*, 21i195-i20:1–5, 1999.

[37] J. Han, J. Pei, and M. Kamber. *Data Mining, Southeast Asia Edition*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 2006.

[38] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

[39] T. K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.

[40] Ionic: Advanced html5 hybrid mobile app framework. http://ionicframework.com. (Accessed on 31/03/2016).

[41] Jersey. https://jersey.java.net. (Accessed on 04/04/2016).

[42] Johns Hopkins Colon Cancer Center. http://www.hopkinscoloncancercenter.org, 2016. (Accessed on 22/02/2016).

[43] M. Katz. *Evaluating clinical and public health interventions : a practical guide to study design and statistics*. New York Cambridge University Press, Cambridge, 2010.

[44] M. Kearns. *Advances in neural information processing systems 11 : proceedings of the 1998 conference*. MIT Press, Cambridge, Mass. London, 1999.

[45] Project kenai. https://jax-rs-spec.java.net. (Accessed on 04/04/2016).

[46] J. Kittler. Combining classifiers: A theoretical framework, 1998.

[47] G. Klepac, G. Klepac, and et al. *Developing Churn Models Using Data Mining Techniques and Social Network Analysis*. IGI Global, Hershey, PA, USA, 1st edition, 2014.

[48] V. Kotu and B. Deshpande. *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*. Elsevier Science, 2014.

[49] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17, 2015.

[50] M. R. C. D. Leon and E. R. L. Jalao. Prediction Model Framework for Imbalanced Datasets. (c):33–41, 2014.

[51] B. Levin, D. A. Lieberman, B. McFarland, R. A. Smith, D. Brooks, K. S. Andrews, C. Dash, F. M. Giardiello, S. Glick, T. R. Levin, P. Pickhardt, D. K. Rex, A. Thorson, and S. J. Winawer. Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: A joint guideline from the american cancer society, the us multi-society task force on colorectal cancer, and the american college of radiology*†. *CA: A Cancer Journal for Clinicians*, 58(3):130–160, 2008.

[52] L. Lima, P. Novais, J. Neves, C. J. Bulas, and R. Costa. Group Decision Making and Quality-of-Information in e-Health Systems. *Logic Journal of the IGPL*, 19(2):315–332, 2011.

[53] R. K. Lindsay, B. G. Buchanan, E. A. Feigenbaum, and J. Lederberg. Applications of artificial intelligence for organic chemistry: the dendral project. *New York*, 1980.

[54] G. Mackenzie. *Statistical modelling in biostatistics and bioinformatics : selected papers*. Springer Verlag, Cham, 2014.

[55] Maven – welcome to apache maven. https://maven.apache.org. (Accessed on 04/04/2016).

[56] Mongodb for giant ideas — mongodb. https://www.mongodb.com. (Accessed on 04/04/2016).

[57] National Cancer Institute. http://www.cancer.gov, 2016. (Accessed on 22/02/2016).

[58] P. Nee. *The Key Facts on Coping With Cancer & Cancer Resources: Everything You Need to Know About Coping With Cancer & Cancer Resources*. The Key Facts on Cancer. Createspace Independent Pub, 2013.

[59] P. Novais, T. Oliveira, and J. Neves. Moving towards a new paradigm of creation, dissemination, and application of computer-interpretable medical knowledge. *Progress in Artificial Intelligence*, 5(2):77–83, 2016.

[60] M. M. Oken, R. H. Creech, D. C. Tormey, J. Horton, T. E. Davis, E. T. McFadden, and P. P. Carbone. Toxicity and response criteria of the eastern cooperative oncology group. *Am J Clin Oncol*, 5(6):649–655, Dec 1982.

[61] T. Oliveira, P. Leão, P. Novais, and J. Neves. Webifying the Computerized Execution of Clinical Practice Guidelines. In J. Bajo Perez, J. M. Corchado Rodríguez, P. Mathieu, A. Campbell, A. Ortega, E. Adam, E. M. Navarro, S. Ahrndt, M. N. Moreno, and V. Julián, editors, *Trends in Practical Applications of Heterogeneous Multi-Agent Systems. The PAAMS Collection SE - 18*, volume 293 of *Advances in Intelligent Systems and Computing*, pages 149–156. Springer International Publishing, 2014.

[62] T. Oliveira, P. Novais, and J. Neves. Development and implementation of clinical guidelines: An artificial intelligence perspective. *Artificial Intelligence Review*, pages 1–29, 2013.

[63] T. Oliveira, P. Novais, and J. Neves. Representation of Clinical Practice Guideline Components in OWL. In J. B. Pérez, R. Hermoso, M. N. Moreno, J. M. C. Rodríguez, B. Hirsch, P. Mathieu, A. Campbell, M. C. Suarez-Figueroa, A. Ortega, E. Adam, and E. Navarro, editors, *Trends in Practical Applications of Agents and Multiagent Systems SE - 10*, volume 221 of *Advances in Intelligent Systems and Computing*, pages 77–85. Springer International Publishing, 2013.

[64] T. Oliveira, K. Satoh, J. Neves, and P. Novais. Applying Speculative Computation to Guideline-Based Decision Support Systems. In *IEEE 27th International Symposium on Computer-Based Medical Systems 2014 (CBMS)*, pages 42–47, 2014.

[65] R. O'Brien. Um exame da abordagem metodológica da pesquisa ação [An Overview of the Methodological Approach of Action Research]. In *Teoria e Prática da Pesquisa Ação [Theory and Practice of Action Research]*, pages 1–15. 2001.

[66] V. L. Patel, E. H. Shortliffe, M. Stefanelli, P. Szolovits, M. R. Berthold, R. Bellazzi, and A. Abu-Hanna. The coming of age of artificial intelligence in medicine. *Artificial Intelligence in Medicine*, 46(1):5 – 17, 2009. Artificial Intelligence in Medicine AIME' 07.

[67] T. S. Portal. Smartphone users worldwide 2014-2019. http://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide. (Accessed on 30/03/2016).

[68] D. M. W. Powers. What the F-measure doesn't measure .... Technical report, Beijing University of Technology, China & Flinders University, Australia.

[69] J. R. Quinlan. *C4.5: Programs for Machine Learning*, volume 1. 1993.

[70] S. Radhakrishnan and D. S. Priyaa. An ensemble approach on missing value handling in hepatitis disease dataset. *International Journal of Computer Applications*, 130(17), 2015.

[71] R. Raj and S. B. Tolety. A study on approaches to build cross-platform mobile applications and criteria to select appropriate approach. In *India Conference (INDICON), 2012 Annual IEEE*, pages 625–629. IEEE, 2012.

[72] RapidMiner. Rapidminer documentation: Aggregate. http://docs.rapidminer.com/studio/operators/data_transformation/aggregation/aggregate.html, 2016. (Accessed on 20/03/2016).

[73] RapidMiner. Rapidminer documentation: Apply model. http://docs.rapidminer.com/studio/operators/scoring/apply_model.html, 2016. (Accessed on 16/05/2016).

[74] RapidMiner. Rapidminer documentation: Bayesian boosting. http://docs.rapidminer.com/studio/operators/modeling/classification_and_regression/meta/bayesian_boosting.html, 2016. (Accessed on 03/01/2016).

[75] RapidMiner. Rapidminer documentation: Filter examples. http://docs.rapidminer.com/studio/operators/data_transformation/filtering/filter_examples.html, 2016. (Accessed on 20/03/2016).

[76] RapidMiner. Rapidminer documentation: Generate attributes. http://docs.rapidminer.com/studio/operators/blending/attributes/generation/generate_attributes.html, 2016. (Accessed on 20/03/2016).

[77] RapidMiner. Rapidminer documentation: Numerical to polynominal. http://docs.rapidminer.com/studio/operators/data_transformation/type_conversion/numerical_to_polynominal.html, 2016. (Accessed on 20/03/2016).

[78] RapidMiner. Rapidminer documentation: Operator reference guide. http://docs.rapidminer.com/studio/operators, 2016. (Accessed on 03/01/2016).

[79] RapidMiner. Rapidminer documentation: Optimize selection. http://docs.rapidminer.com/studio/operators/data_transformation/attribute_space_transformation/selection/optimization/optimize_selection.html, 2016. (Accessed on 03/01/2016).

[80] RapidMiner. Rapidminer documentation: Replace missing values. http://docs.rapidminer.com/studio/operators/data_transformation/data_cleansing/replace_missing_values.html, 2016. (Accessed on 20/03/2016).

[81] RapidMiner. Rapidminer documentation: Sample. http://docs.rapidminer.com/studio/operators/blending/examples/sampling/sample.html, 2016. (Accessed on 29/05/2016).

[82] RapidMiner. Rapidminer documentation: Sample (bootstrapping). http://docs.rapidminer.com/studio/operators/blending/examples/sampling/sample_bootstrapping.html, 2016. (Accessed on 29/05/2016).

[83] RapidMiner. Rapidminer documentation: Select attributes. http://docs.rapidminer.com/studio/operators/data_transformation/attribute_space_transformation/selection/select_attributes.html, 2016. (Accessed on 20/03/2016).

[84] P. Refaeilzadeh, L. Tang, and H. Liu. Cross-validation. In L. LIU and M. ÖZSU, editors, *Encyclopedia of Database Systems*, pages 532–538. Springer US, 2009.

[85] L. a. Renfro, A. Grothey, Y. Xue, L. B. Saltz, T. André, C. Twelves, R. Labianca, C. J. Allegra, S. R. Alberts, C. L. Loprinzi, G. Yothers, and D. J. Sargent. Cancer prediction tools: Stage iii colon cancer. http://www.mayoclinic.org/medical-professionals/cancer-prediction-tools/colon-cancer. (Accessed on 26/03/2016).

[86] L. a. Renfro, A. Grothey, Y. Xue, L. B. Saltz, T. André, C. Twelves, R. Labianca, C. J. Allegra, S. R. Alberts, C. L. Loprinzi, G. Yothers, and D. J. Sargent. ACCENT-Based

Web Calculators to Predict Recurrence and Overall Survival in Stage III Colon Cancer. *Journal of the National Cancer Institute*, 106(12):1–9, 2014.

[87] S. SHERER. An introduction to radiation protection 6e. 2002.

[88] K. Siau and Z. Shen. Mobile healthcare informatics. *Medical Informatics and the Internet in Medicine*, 31(2):89–99, 2006.

[89] P. B. Snow, D. J. Kerr, J. M. Brandt, and D. M. Rodvold. Neural network and regression predictions of 5-year survival after colon carcinoma treatment. *Cancer*, 91(8 Suppl):1673–1678, 2001.

[90] A. C. Society. Colorectal cancer facts & figures 2014-2016. *Colorectal Cancer Facts and Figures*, pages 1–32, 2014.

[91] S. Steele, A. Bilchik, J. Eberhardt, P. Kalina, A. Nissan, E. Johnson, I. Avital, and A. Stojadinovic. Using machine-learned bayesian belief networks to predict perioperative risk of clostridium difficile infection following colon surgery. *Interactive journal of medical research*, 1(2):e6, Jan. 2012.

[92] A. Stojadinovic, A. Bilchik, D. Smith, J. S. Eberhardt, E. B. Ward, A. Nissan, E. K. Johnson, M. Protic, G. E. Peoples, I. Avital, and S. R. Steele. Clinical Decision Support and Individualized Prediction of Survival in Colon Cancer: Bayesian Belief Network Model. *Annals of Surgical Oncology*, 20(1):161–174, 2012.

[93] P. Szolovits. *Artificial Intelligence and Medicine*. Westview Press, Boulder, Colorado, 1982.

[94] T. Taniyama, K. Hashimoto, N. Katsumata, A. Hirakawa, K. Yonemori, M. Yunokawa, C. Shimizu, K. Tamura, M. Ando, and Y. Fujiwara. Can oncologists predict survival for patients with progressive disease after standard chemotherapies? *Current Oncology*, 21(2):84, 2014.

[95] N. C. I. U. S. National Institutes of Health. Seer training modules, colorectal cancer. http://training.seer.cancer.gov. (Accessed on 27/12/2015).

[96] S. Unnikrishnan, S. Surve, and D. Bhoir. *Advances in Computing, Communication and Control: International Conference, ICAC3 2011, Mumbai, India, January 28-29, 2011. Proceedings*. Communications in Computer and Information Science. Springer Berlin Heidelberg, 2011.

[97] V. Valentini, R. G. van Stiphout, G. Lammering, M. A. Gambacorta, M. C. Barba, M. Bebenek, F. Bonnetain, J.-F. Bosset, K. Bujko, L. Cionini, J.-P. Gerard, C. Rödel, A. Sainato, R. Sauer, B. D. Minsky, L. Collette, and P. Lambin. Rectal carcinoma: local

and distant control and survival. http://www.predictcancer.org/Main.php?page=RectumFollowUpModel. (Accessed on 26/03/2016).

[98] V. Valentini, R. G. van Stiphout, G. Lammering, M. A. Gambacorta, M. C. Barba, M. Bebenek, F. Bonnetain, J.-F. Bosset, K. Bujko, L. Cionini, J.-P. Gerard, C. Rödel, A. Sainato, R. Sauer, B. D. Minsky, L. Collette, and P. Lambin. Nomograms for Predicting Local Recurrence, Distant Metastases, and Overall Survival for Patients With Locally Advanced Rectal Cancer on the Basis of European Randomized Clinical Trials. *Journal of Clinical Oncology*, 29(23):3163 –3172, 2011.

[99] T. van der Ploeg, F. Datema, R. B. de Jong, and E. W. Steyerberg. Prediction of survival with alternative modeling techniques using pseudo values. *PloS one*, 9(6):e100234, 2014.

[100] C. L. Ventola. Mobile devices and apps for health care professionals: uses and benefits. *Pharmacy and Therapeutics*, 39(5):356, 2014.

[101] S. J. Wang, A. R. Wissel, J. Y. Luh, C. D. Fuller, J. Kalpathy-Cramer, and C. R. Thomas. Cancer survival prediction calculators: Rectal cancer. http://skynet.ohsu.edu/nomograms/gastrointestinal/rectal.php. (Accessed on 26/03/2016).

[102] S. J. Wang, A. R. Wissel, J. Y. Luh, C. D. Fuller, J. Kalpathy-Cramer, and C. R. Thomas. An interactive tool for individualized estimation of conditional survival in rectal cancer. *Annals of surgical oncology*, 18(6):1547–52, 2011.

[103] M. R. Weiser, M. Gönen, J. F. Chou, M. W. Kattan, and D. Schrag. Overall survival probability following surgery. https://www.mskcc.org/nomograms/colorectal/overall-survival-probability. (Accessed on 26/03/2016).

[104] M. R. Weiser, M. Gönen, J. F. Chou, M. W. Kattan, and D. Schrag. Predicting survival after curative colectomy for cancer: Individualizing colon cancer staging. *Journal of Clinical Oncology*, 29(36):4796–4802, 2011.

[105] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. 2005.

[106] A. C. Wolff, M. E. H. Hammond, J. N. Schwartz, and et al. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. *Journal of clinical oncology*, 25(1):18–43, 2007.

[107] H. Zhao and P. Doshi. Towards automated restful web service composition. In *Web Services, 2009. ICWS 2009. IEEE International Conference on*, pages 189–196. IEEE, 2009.

# A

## SCRIPT TO PROCESS THE SEER DATASET

In this appendix is presented the script used to convert the raw data from SEER database into a csv file.

### A.1 C# CODE

```
using System;
using System.Collections.Generic;
using System.IO;
using System.Linq;
using System.Text;
using System.Threading.Tasks;

namespace seerdata2csv
{
    class Program
    {
        static void Main(string[] args)
        {
            SEER2CSV();
        }

        public static void SEER2CSV()
        {
            try
            {
string[] lines = File.ReadAllLines(@"//seer/COLRECT.TXT");
                StreamWriter sw = new StreamWriter("seer_data.csv", true);
```

**Appendix A. script to process the seer dataset**

```
sw.WriteLine("Patient ID number;
Registry ID;
Marital Status at DX;
Race/Ethnicity;
Spanish/Hispanic Origin;
NHIA Derived Hispanic Origin;
Sex;
Age at diagnosis;
Year of Birth;
Sequence Number|Central;
Month of diagnosis;
Year of diagnosis;
Primary Site;
Laterality;
Histology (92-00) ICD-O-2;Behavior (92-00) ICD-O-2;
Histologic Type ICD-O-3;
Behavior Code ICD-O-3;
Grade;Diagnostic Confirmation;
Type of Reporting Source;
EOD|Tumor Size;
EOD|Extension;
EOD|Extension Prost Path;
EOD|Lymph Node Involv;
Regional Nodes Positive;
Regional Nodes Examined;
EOD|Old 13;
Digit EOD|Old 2;
Digit EOD|Old 4 Digit;
Coding System for EOD;
Tumor Marker 1;
Tumor Marker 2;
Tumor Marker 3;
CS Tumor Size;
CS Extension;
CS Lymph Nodes;
CS Mets at Dx;
CS Site-Specific Factor 1;
CS Site-Specific Factor 2;
```

```
CS Site-Specific Factor 3;
CS Site-Specific Factor 4;
CS Site-Specific Factor 5;
CS Site-Specific Factor 6;
CS Site-Specific Factor 25;
Derived AJCC T;
Derived AJCC N;
Derived AJCC M;
Derived AJCC Stage Group;
Derived SS1977;
Derived SS2000;
Derived AJCC|Flag;
Derived SS1977|Flag;
Derived SS2000|Flag;
CS Version Input Original;
CS Version Derived;
CS Version Input Current;
RX Summ|Surg Prim Site;
RX Summ|Scope Reg LN Sur;
RX Summ|Surg Oth Reg/Dis;
RX Summ|Reg LN Examined;
Reason for no surgery;
RX Summ|Radiation;
RX Summ|Rad to CNS;
RX Summ|Surg / Rad Seq;
RX Summ|Surgery Type;
RX Summ|Scope Reg 98-02;
RX Summ|Surg Oth 98-02;
SEER Record Number;
Over-ride age/site/morph;
Over-ride seqno/dxconf;
Over-ride site/lat/seqno;
Over-ride surg/dxconf;
Over-ride site/type;
Over-ride histology;
Over-ride report source;
Over-ride ill-define site;
Over-ride Leuk, Lymph;
```

**Appendix A. script to process the seer dataset**

```
Over-ride site/behavior;
Over-ride site/eod/dx dt;
Over-ride site/lat/eod;
Over-ride site/lat/morph;
SEER Type of Follow-up;
Age Recode <1 Year olds;
Site Recode ICD-O-3/WHO 2008;
Recode ICD-O-2 to 9;
Recode ICD-O-2 to 10;
ICCC site recode ICD-O-3/WHO 2008;
ICCC site rec extended ICD-O-3/WHO 2008;
Behavior Recode for Analysis;
Histology Recode|Broad Groupings;
Histology Recode|Brain Groupings;
CS Schema v0204+;
Race recode (White, Black, Other);
Race recode (W, B, AI, API);
Origin recode NHIA (Hispanic, Non-Hisp);
SEER historic stage A;
AJCC stage 3rd edition (1988-2003);
SEER modified AJCC Stage 3rd ed (1988- 2003);
SEER Summary Stage 1977 (1995-2000);
SEER Summary Stage 2000 (2001-2003);
Number of primaries;
First malignant primary indicator;
State-county recode;
Cause of Death to SEER site recode;
COD to site rec KM;Vital Status recode;
IHS Link;
Summary stage 2000 (1998+);
AYA site recode/WHO 2008;
Lymphoma subtype recode/WHO 2008;
SEER Cause-Specific Death Classification;
SEER Other Cause of Death Classification;
CS Tumor Size/Ext Eval;
CS Lymph Nodes Eval;
CS Mets Eval;
Primary by international rules;
```

```
ER Status Recode Breast Cancer (1990+);
PR Status Recode Breast Cancer (1990+);
CS Schema -AJCC 6th ed (previously called v1);
CS Site-Specific Factor 8;
CS Site-Specific Factor 10;
CS Site-Specific Factor 11;
CS Site-Specific Factor 13;
CS Site-Specific Factor 15;
CS Site-Specific Factor 16;
Lymph vascular invasion;
Survival months;
Survival months flag;
Survival months { presumed alive;
Survival months flag { presumed alive;
Insurance recode (2007+);
Derived AJCC-7 T;
Derived AJCC-7 N;
Derived AJCC-7 M;
Derived AJCC-7 Stage Grp;
Breast Adjusted AJCC 6th T (1988+);
Breast Adjusted AJCC 6th N (1988+);
Breast Adjusted AJCC 6th M (1988+);
Breast Adjusted AJCC 6th Stage (1988+);
CS Site-Specific Factor 7;
CS Site-Specific Factor 9;
CS Site-Specific Factor 12;
Derived HER2 Recode (2010+);
Breast Subtype (2010+);
Lymphomas: Ann Arbor Staging (1983+)");
                foreach (string line in lines)
                {
                    sw.WriteLine(line.Substring(0, 8).Trim() + ';' +
                     line.Substring(17, 10).Trim() + ';' +
                     line.Substring(18, 1).Trim() + ';' +
                     line.Substring(19, 2).Trim() + ';' +
                     line.Substring(21, 1).Trim() + ';' +
                     line.Substring(22, 1).Trim() + ';' +
                     line.Substring(23, 1).Trim() + ';' +
```

```
                    line.Substring(24, 3).Trim() + ';' +
                    line.Substring(27, 4).Trim() + ';' +
                    line.Substring(34, 2).Trim() + ';' +
                    line.Substring(36, 2).Trim() + ';' +
                    line.Substring(38, 4).Trim() + ';' +
                    line.Substring(42, 4).Trim() + ';' +
                    line.Substring(46, 1).Trim() + ';' +
                    line.Substring(47, 4).Trim() + ';' +
                    line.Substring(51, 1).Trim() + ';' +
line.Substring(52, 4).Trim() + ';' +
line.Substring(56, 1).Trim() + ';' +
line.Substring(57, 1).Trim() + ';' +
line.Substring(58, 1).Trim() + ';' +
line.Substring(59, 1).Trim() + ';' +
line.Substring(60, 3).Trim() + ';' +
line.Substring(63, 2).Trim() + ';' +
line.Substring(65, 2).Trim() + ';' +
line.Substring(67, 1).Trim() + ';' +
line.Substring(68, 2).Trim() + ';' +
line.Substring(70, 2).Trim() + ';' +
line.Substring(72, 13).Trim() + ';' +
line.Substring(85, 2).Trim() + ';' +
line.Substring(87, 4).Trim() + ';' +
line.Substring(91, 1).Trim() + ';' +
line.Substring(92, 1).Trim() + ';' +
line.Substring(93, 1).Trim() + ';' +
line.Substring(94, 1).Trim() + ';' +
line.Substring(95, 3).Trim() + ';' +
line.Substring(98, 3).Trim() + ';' +
line.Substring(101, 3).Trim() + ';' +
line.Substring(104, 2).Trim() + ';' +
line.Substring(106, 3).Trim() + ';' +
line.Substring(109, 3).Trim()+ ';' +
line.Substring(112, 3).Trim() + ';' +
line.Substring(115, 3).Trim() + ';' +
line.Substring(118, 3).Trim() + ';' +
line.Substring(121, 3).Trim() + ';' +
line.Substring(124, 3).Trim() + ';' +
```

```
line.Substring(127, 2).Trim() + ';' +
line.Substring(129, 2).Trim() + ';' +
line.Substring(131, 2).Trim() + ';' +
line.Substring(133, 2).Trim() + ';' +
line.Substring(135, 1).Trim() + ';' +
line.Substring(136, 1).Trim() + ';' +
line.Substring(137, 1).Trim() + ';' +
line.Substring(138, 1).Trim() + ';' +
line.Substring(139, 1).Trim() + ';' +
line.Substring(140, 6).Trim() + ';' +
line.Substring(146, 6).Trim() + ';' +
line.Substring(152, 6).Trim() + ';' +
line.Substring(158, 2).Trim() + ';' +
line.Substring(160, 1).Trim() + ';' +
line.Substring(161, 1).Trim() + ';' +
line.Substring(162, 2).Trim() + ';' +
line.Substring(165, 1).Trim() + ';' +
line.Substring(166, 1).Trim() + ';' +
line.Substring(167, 1).Trim() + ';' +
line.Substring(168, 1).Trim() + ';' +
line.Substring(169, 2).Trim() + ';' +
line.Substring(173, 1).Trim() + ';' +
line.Substring(174, 1).Trim() + ';' +
line.Substring(175, 2).Trim() + ';' +
line.Substring(177, 1).Trim() + ';' +
line.Substring(178, 1).Trim() + ';' +
line.Substring(179, 1).Trim() + ';' +
line.Substring(180, 1).Trim() + ';' +
line.Substring(181, 1).Trim() + ';' +
line.Substring(182, 1).Trim() + ';' +
line.Substring(183, 1).Trim() + ';' +
line.Substring(184, 1).Trim() + ';' +
line.Substring(185, 1).Trim() + ';' +
line.Substring(186, 1).Trim() + ';' +
line.Substring(187, 1).Trim() + ';' +
line.Substring(188, 1).Trim() + ';' +
line.Substring(189, 1).Trim() + ';' +
line.Substring(190, 1).Trim() + ';' +
```

```
line.Substring(191, 2).Trim() + ';' +
line.Substring(198, 5).Trim() + ';' +
line.Substring(203, 4).Trim() + ';' +
line.Substring(207, 4).Trim() + ';' +
line.Substring(217, 3).Trim() + ';' +
line.Substring(220, 3).Trim() + ';' +
line.Substring(223, 1).Trim() + ';' +
line.Substring(225, 2).Trim() + ';' +
line.Substring(227, 2).Trim() + ';' +
line.Substring(229, 3).Trim() + ';' +
line.Substring(232, 1).Trim() + ';' +
line.Substring(233, 1).Trim() + ';' +
line.Substring(234, 1).Trim() + ';' +
line.Substring(235, 1).Trim() + ';' +
line.Substring(236, 2).Trim() + ';' +
line.Substring(238, 2).Trim() + ';' +
line.Substring(240, 1).Trim() + ';' +
line.Substring(241, 1).Trim() + ';' +
line.Substring(242, 2).Trim() + ';' +
line.Substring(244, 1).Trim() + ';' +
line.Substring(245, 5).Trim() + ';' +
line.Substring(254, 5).Trim() + ';' +
line.Substring(259, 5).Trim() + ';' +
line.Substring(264, 1).Trim() + ';' +
line.Substring(265, 1).Trim() + ';' +
line.Substring(266, 1).Trim() + ';' +
line.Substring(267, 2).Trim() + ';' +
line.Substring(269, 2).Trim() + ';' +
line.Substring(271, 1).Trim() + ';' +
line.Substring(272, 1).Trim() + ';' +
line.Substring(273, 1).Trim() + ';' +
line.Substring(274, 1).Trim() + ';' +
line.Substring(275, 1).Trim() + ';' +
line.Substring(276, 1).Trim() + ';' +
line.Substring(277, 1).Trim() + ';' +
line.Substring(278, 1).Trim() + ';' +
line.Substring(279, 2).Trim() + ';' +
line.Substring(281, 3).Trim() + ';' +
```

```
line.Substring(284, 3).Trim() + ';' +
line.Substring(287, 3).Trim() + ';' +
line.Substring(290, 3).Trim() + ';' +
line.Substring(293, 3).Trim() + ';' +
line.Substring(296, 3).Trim() + ';' +
line.Substring(299, 1).Trim() + ';' +
line.Substring(300, 4).Trim() + ';' +
line.Substring(304, 1).Trim() + ';' +
line.Substring(305, 4).Trim() + ';' +
line.Substring(309, 1).Trim() + ';' +
line.Substring(310, 1).Trim() + ';' +
line.Substring(311, 3).Trim() + ';' +
line.Substring(314, 3).Trim() + ';' +
line.Substring(317, 3).Trim() + ';' +
line.Substring(320, 3).Trim() + ';' +
line.Substring(323, 2).Trim() + ';' +
line.Substring(325, 2).Trim() + ';' +
line.Substring(327, 2).Trim() + ';' +
line.Substring(329, 2).Trim() + ';' +
line.Substring(331, 3).Trim() + ';' +
line.Substring(334, 3).Trim() + ';' +
line.Substring(337, 3).Trim() + ';' +
line.Substring(340, 1).Trim() + ';' +
line.Substring(341, 1).Trim() + ';' +
line.Substring(347, 1).Trim());
                }
                sw.Flush();
                sw.Close();
            }
            catch (Exception ex)
            {
                Console.WriteLine(ex.ToString());
                Console.ReadLine();
            }
        }
    }
}
```

# B

# RAPIDMINER PROCESSES

In this appendix is presented the workflows constructed in the RapidMiner software to develop the prediction models.

## B.1 PREPROCESSING PROCESS

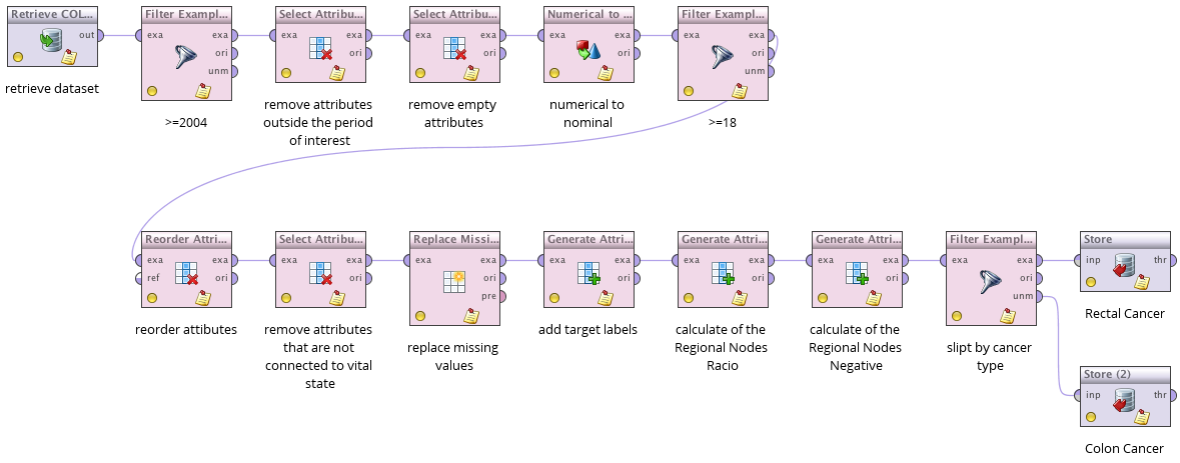



Figure B.1.44.: Preprocessing phase, part 1 of 2.

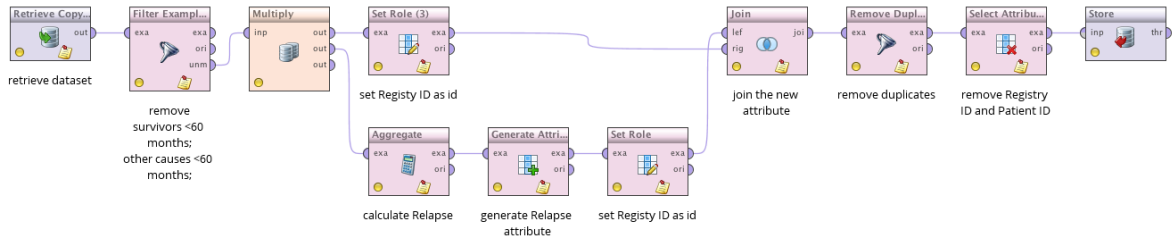



Figure B.1.45.: Preprocessing phase, part 2 of 2.

## B.2 SPLIT DATASET

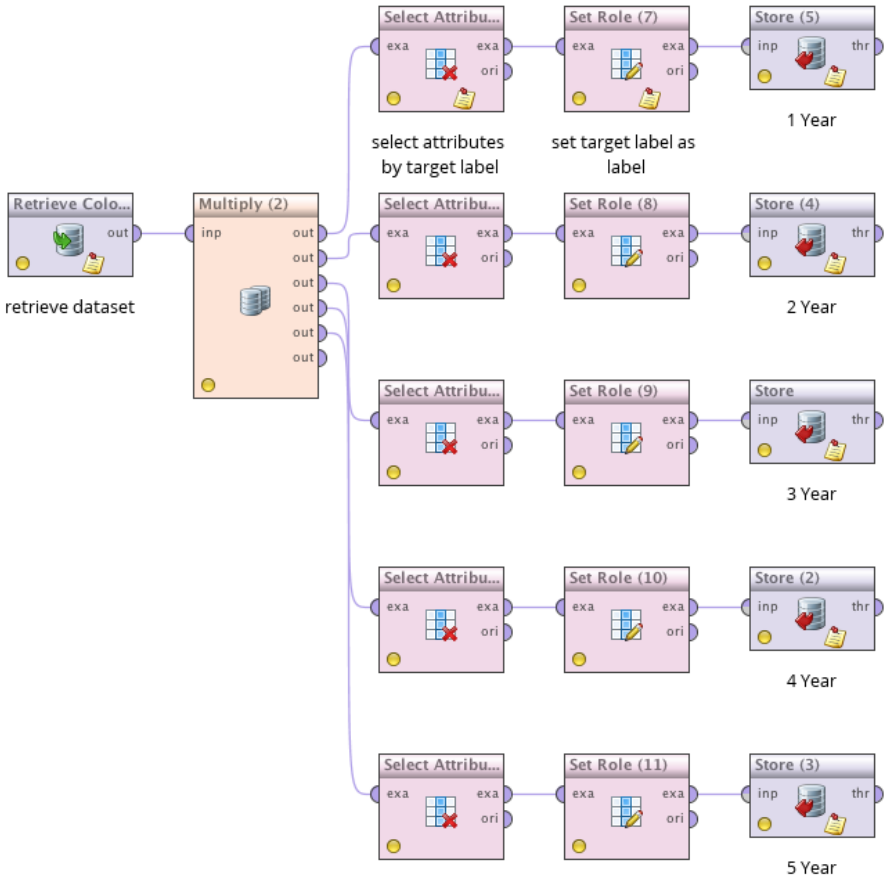



Figure B.2.46.: Split dataset phase.

## B.3 FEATURE SELECTION

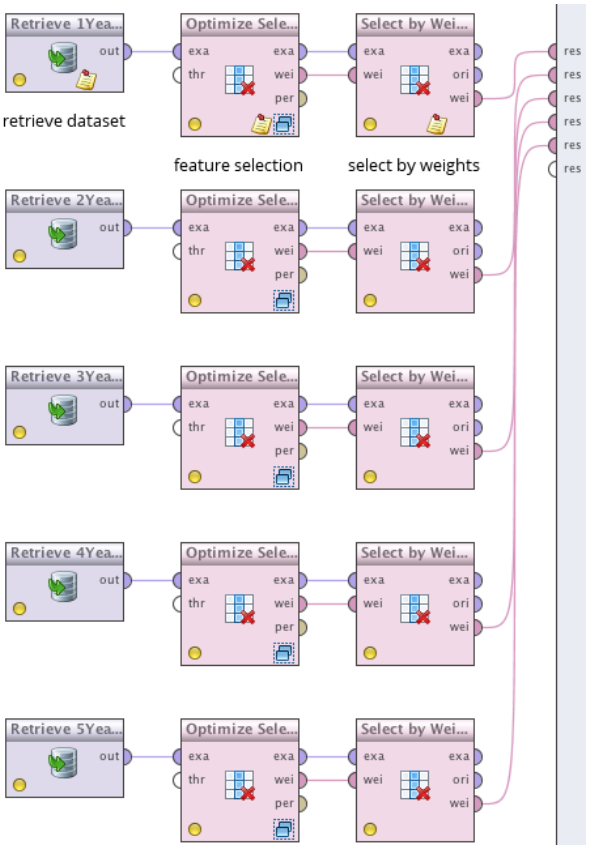



Figure B.3.47.: Feature selection phase.

## B.4 SAMPLING DATA

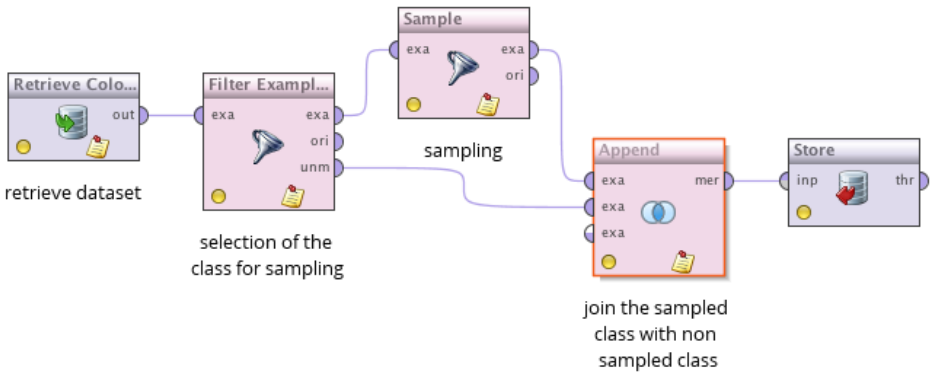



Figure B.4.48.: An example of the sampling data phase.

## B.5 MODELING AND EVALUATION

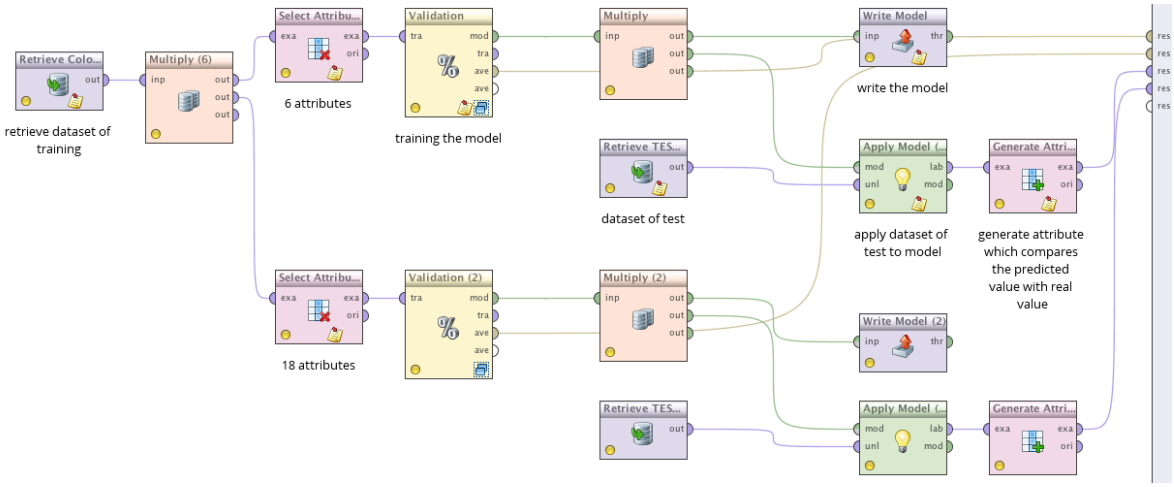



Figure B.5.49.: An example of the modeling and evaluation phase.

# C

# DETAILS OF RESULTS

In this appendix is presented all the details of the results. It was divided by cancer type and dataset type used to train the models.

## Appendix C. details of results

### C.1 SURVIVABILITY PREDICTION MODELS

#### C.1.1 *Colon Cancer*

*Unbalanced Models Without Unknowns*

Table C.1.1.: Survivability percentage accuracy of unbalanced models without unknowns, for colon cancer.

| Accuracy - Colon Cancer Unbalanced Models Without Unknowns | | | | | | |
|---|---|---|---|---|---|---|
| | 1Year | 2Year | 3Year | 4Year | 5Year | Average |
| | 6 attributes | 6 attributes | 6 attributes | 6 attributes | 6 attributes | 6 attributes |
| Stacking | 95,85% | 96,74% | 96,88% | 97,01% | 97,03% | 96,70% |
| Voting | 96,24% | 96,30% | 96,70% | 97,10% | 97,00% | 96,67% |
| Bayesian Boosting with Decision Tree | 95,92% | 96,03% | 96,57% | 96,64% | 96,71% | 80,31% |
| AdaBoost with Decision Tree | 95,99% | 96,09% | 96,67% | 96,67% | 96,72% | 96,43% |
| Bagging with Decision Tree | 95,82% | 95,58% | 95,96% | 95,55% | 95,51% | 95,68% |
| Bayesian Boosting with Random Forest | 88,17% | 88,52% | 88,56% | 88,89% | 89,76% | 88,78% |
| AdaBoost with Random Forest | 84,45% | 84,69% | 86,67% | 87,59% | 88,75% | 86,43% |
| Bagging with Random Forest | 85,72% | 87,92% | 89,37% | 90,08% | 90,14% | 88,65% |
| Bayesian Boosting with Naive Bayes | 84,98% | 84,25% | 85,81% | 85,45% | 85,37% | 85,17% |
| AdaBoost with Naive Bayes | 84,77% | 84,35% | 85,77% | 85,36% | 85,36% | 85,12% |
| Bagging with Naive Bayes | 84,67% | 84,40% | 85,75% | 85,44% | 85,37% | 85,13% |
| Bayesian Boosting with K-NN | 95,11% | 95,38% | 95,67% | 95,69% | 95,91% | 95,55% |
| AdaBoost with K-NN | 95,18% | 95,28% | 95,62% | 95,82% | 96,02% | 95,58% |
| Bagging with K-NN | 95,24% | 95,41% | 95,74% | 95,77% | 95,87% | 95,61% |

Table C.1.2.: Survivability percentage AUC of unbalanced models without unknowns, for colon cancer.

| AUC - Colon Cancer Unbalanced Models Without Unknowns | | | | | | |
|---|---|---|---|---|---|---|
| | 1Year | 2Year | 3Year | 4Year | 5Year | Average |
| | 6 attributes | 6 attributes | 6 attributes | 6 attributes | 6 attributes | 6 attributes |
| Stacking | 0,982 | 0,985 | 0,988 | 0,988 | 0,988 | 0,986 |
| Voting | 0,959 | 0,972 | 0,978 | 0,975 | 0,974 | 0,972 |
| Bayesian Boosting with Decision Tree | 0,935 | 0,948 | 0,958 | 0,968 | 0,967 | 0,796 |
| AdaBoost with Decision Tree | 0,939 | 0,961 | 0,965 | 0,972 | 0,973 | 0,962 |
| Bagging with Decision Tree | 0,945 | 0,949 | 0,946 | 0,941 | 0,939 | 0,944 |
| Bayesian Boosting with Random Forest | 0,833 | 0,892 | 0,909 | 0,924 | 0,927 | 0,897 |
| AdaBoost with Random Forest | 0,798 | 0,877 | 0,902 | 0,921 | 0,920 | 0,884 |
| Bagging with Random Forest | 0,943 | 0,955 | 0,966 | 0,967 | 0,966 | 0,959 |
| Bayesian Boosting with Naive Bayes | 0,875 | 0,888 | 0,891 | 0,907 | 0,908 | 0,894 |
| AdaBoost with Naive Bayes | 0,860 | 0,872 | 0,884 | 0,907 | 0,908 | 0,886 |
| Bagging with Naive Bayes | 0,891 | 0,901 | 0,919 | 0,925 | 0,925 | 0,912 |
| Bayesian Boosting with K-NN | 0,926 | 0,942 | 0,952 | 0,954 | 0,957 | 0,946 |
| AdaBoost with K-NN | 0,927 | 0,941 | 0,951 | 0,956 | 0,958 | 0,947 |
| Bagging with K-NN | 0,947 | 0,958 | 0,965 | 0,967 | 0,967 | 0,961 |

Table C.1.3.: F-measure performance of unbalanced models without unknowns, for colon cancer.

| F-Measure - Colon Cancer Unbalanced Models Without Unknowns | | | | | | |
|---|---|---|---|---|---|---|
| | 1Year | 2Year | 3Year | 4Year | 5Year | Average |
| | 6 attributes | 6 attributes | 6 attributes | 6 attributes | 6 attributes | 6 attributes |
| Stacking | 90,45% | 94,63% | 95,66% | 97,56% | 97,50% | 95,16% |
| Voting | 90,86% | 93,77% | 95,38% | 97,63% | 97,49% | 95,03% |
| Bayesian Boosting with Decision Tree | 90,36% | 93,43% | 95,21% | 97,24% | 97,22% | 78,91% |
| AdaBoost with Decision Tree | 90,51% | 93,53% | 95,36% | 97,26% | 97,23% | 94,78% |
| Bagging with Decision Tree | 89,87% | 92,49% | 94,26% | 96,29% | 96,15% | 93,81% |
| Bayesian Boosting with Random Forest | 66,64% | 78,96% | 82,43% | 91,28% | 91,69% | 82,20% |
| AdaBoost with Random Forest | 46,57% | 69,38% | 79,57% | 90,38% | 90,95% | 75,37% |
| Bagging with Random Forest | 51,09% | 76,87% | 83,63% | 92,25% | 92,11% | 79,19% |
| Bayesian Boosting with Naive Bayes | 61,07% | 72,58% | 79,52% | 88,38% | 88,07% | 77,93% |
| AdaBoost with Naive Bayes | 61,95% | 73,30% | 79,46% | 88,31% | 88,07% | 78,22% |
| Bagging with Naive Bayes | 61,71% | 73,39% | 79,43% | 88,39% | 88,07% | 78,20% |
| Bayesian Boosting with K-NN | 88,69% | 92,13% | 93,90% | 96,58% | 96,66% | 93,60% |
| AdaBoost with K-NN | 88,69% | 92,13% | 93,90% | 96,58% | 96,66% | 93,60% |
| Bagging with K-NN | 88,87% | 92,37% | 94,07% | 96,54% | 96,54% | 93,68% |

Table C.1.4.: Percentage of wrongly classified cases of unbalanced models without unknowns, for colon cancer.

| Wrongly Classified Cases - Colon Cancer Unbalanced Models Without Unknowns | | | | | | |
|---|---|---|---|---|---|---|
| | 1Year | 2Year | 3Year | 4Year | 5Year | Average |
| | 6 attributes | 6 attributes | 6 attributes | 6 attributes | 6 attributes | 6 attributes |
| Stacking | 3,29% | 2,70% | 2,88% | 2,84% | 2,84% | 2,91% |
| Voting | 3,02% | 2,97% | 3,38% | 3,06% | 3,02% | 3,09% |
| Bayesian Boosting with Decision Tree | 3,33% | 3,38% | 3,69% | 3,33% | 3,51% | 3,45% |
| AdaBoost with Decision Tree | 3,38% | 3,33% | 3,65% | 3,33% | 3,47% | 3,43% |
| Bagging with Decision Tree | 3,51% | 3,92% | 4,41% | 4,32% | 3,83% | 4,00% |
| Bayesian Boosting with Random Forest | 10,81% | 13,15% | 12,56% | 10,04% | 9,64% | 11,24% |
| AdaBoost with Random Forest | 14,99% | 17,15% | 17,15% | 14,41% | 13,51% | 15,44% |
| Bagging with Random Forest | 15,49% | 12,25% | 10,67% | 10,72% | 10,04% | 11,83% |
| Bayesian Boosting with Naive Bayes | 14,77% | 16,25% | 14,50% | 14,00% | 14,45% | 12,33% |
| AdaBoost with Naive Bayes | 15,35% | 16,48% | 14,50% | 14,00% | 14,45% | 14,96% |
| Bagging with Naive Bayes | 15,35% | 16,48% | 14,41% | 14,05% | 14,32% | 14,92% |
| Bayesian Boosting with K-NN | 3,83% | 4,14% | 3,74% | 3,60% | 3,69% | 3,80% |
| AdaBoost with K-NN | 3,83% | 4,14% | 3,74% | 3,60% | 3,69% | 3,80% |
| Bagging with K-NN | 3,83% | 4,14% | 3,74% | 3,60% | 3,69% | 3,80% |

## Appendix C. details of results

*Unbalanced Models*

Table C.1.5.: Survivability percentage accuracy of unbalanced models, for colon cancer.

| Accuracy - Colon Cancer Unbalanced Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 96,88% | 95,66% | 96,98% | 96,20% | 97,16% | 96,44% | 97,28% | 96,69% | 97,23% | 96,45% | 97,11% | 96,29% |
| Voting | 97,07% | 95,88% | 96,91% | 96,00% | 97,05% | 96,37% | 97,21% | 96,67% | 97,19% | 96,32% | 97,09% | 96,25% |
| Bayesian Boosting with Decision Tree | 96,55% | 95,76% | 96,48% | 95,74% | 96,74% | 96,19% | 96,87% | 96,24% | 96,69% | 96,06% | 80,56% | 80,00% |
| AdaBoost with Decision Tree | 96,55% | 95,78% | 96,48% | 95,81% | 96,74% | 96,28% | 96,87% | 96,27% | 96,69% | 96,09% | 96,67% | 96,05% |
| Bagging with Decision Tree | 96,42% | 95,48% | 96,14% | 95,17% | 96,20% | 95,53% | 96,09% | 95,16% | 96,21% | 94,85% | 96,21% | 95,24% |
| Bayesian Boosting with Random Forest | 86,20% | 88,00% | 85,94% | 87,94% | 85,64% | 88,83% | 85,64% | 89,45% | 85,38% | 89,46% | 85,76% | 88,74% |
| AdaBoost with Random Forest | 83,08% | 86,23% | 82,68% | 86,33% | 83,64% | 84,62% | 84,58% | 88,87% | 84,25% | 88,49% | 83,65% | 86,91% |
| Bagging with Random Forest | 83,51% | 87,13% | 83,88% | 88,20% | 85,03% | 88,92% | 85,54% | 89,75% | 85,49% | 89,49% | 84,69% | 88,70% |
| Bayesian Boosting with Naive Bayes | 83,54% | 82,77% | 84,28% | 83,40% | 85,58% | 84,71% | 84,38% | 85,54% | 84,17% | 85,81% | 84,39% | 84,45% |
| AdaBoost with Naive Bayes | 84,86% | 82,70% | 84,76% | 83,61% | 84,80% | 84,68% | 84,40% | 85,51% | 84,86% | 85,83% | 84,74% | 84,47% |
| Bagging with Naive Bayes | 82,17% | 82,43% | 81,61% | 83,58% | 80,33% | 84,69% | 79,88% | 85,54% | 79,73% | 85,87% | 80,74% | 84,42% |
| Bayesian Boosting with K-NN | 96,52% | 94,89% | 96,15% | 94,98% | 95,99% | 95,20% | 96,13% | 95,34% | 95,98% | 95,23% | 96,15% | 95,13% |
| AdaBoost with K-NN | 96,52% | 94,89% | 96,20% | 95,00% | 96,06% | 95,25% | 96,03% | 95,36% | 95,92% | 95,19% | 96,15% | 95,14% |
| Bagging with K-NN | 96,52% | 94,82% | 96,18% | 94,89% | 96,05% | 95,30% | 96,03% | 95,38% | 96,04% | 95,20% | 96,16% | 95,12% |

Table C.1.6.: Survivability percentage AUC of unbalanced models, for colon cancer.

| AUC - Colon Cancer Unbalanced Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 0,985 | 0,980 | 0,989 | 0,984 | 0,990 | 0,986 | 0,990 | 0,988 | 0,990 | 0,985 | 0,989 | 0,985 |
| Voting | 0,973 | 0,960 | 0,980 | 0,969 | 0,982 | 0,974 | 0,978 | 0,969 | 0,978 | 0,966 | 0,978 | 0,968 |
| Bayesian Boosting with Decision Tree | 0,951 | 0,939 | 0,961 | 0,945 | 0,972 | 0,960 | 0,966 | 0,966 | 0,978 | 0,959 | 0,805 | 0,795 |
| AdaBoost with Decision Tree | 0,967 | 0,944 | 0,966 | 0,954 | 0,977 | 0,962 | 0,973 | 0,964 | 0,976 | 0,963 | 0,972 | 0,957 |
| Bagging with Decision Tree | 0,953 | 0,940 | 0,959 | 0,945 | 0,959 | 0,950 | 0,944 | 0,943 | 0,948 | 0,938 | 0,953 | 0,943 |
| Bayesian Boosting with Random Forest | 0,788 | 0,865 | 0,855 | 0,901 | 0,864 | 0,916 | 0,881 | 0,930 | 0,891 | 0,926 | 0,856 | 0,908 |
| AdaBoost with Random Forest | 0,755 | 0,858 | 0,810 | 0,893 | 0,842 | 0,888 | 0,864 | 0,921 | 0,866 | 0,921 | 0,827 | 0,896 |
| Bagging with Random Forest | 0,938 | 0,949 | 0,932 | 0,957 | 0,937 | 0,961 | 0,937 | 0,965 | 0,939 | 0,960 | 0,937 | 0,958 |
| Bayesian Boosting with Naive Bayes | 0,888 | 0,876 | 0,904 | 0,893 | 0,910 | 0,908 | 0,910 | 0,915 | 0,908 | 0,912 | 0,904 | 0,901 |
| AdaBoost with Naive Bayes | 0,883 | 0,855 | 0,901 | 0,882 | 0,908 | 0,898 | 0,914 | 0,913 | 0,914 | 0,911 | 0,904 | 0,892 |
| Bagging with Naive Bayes | 0,872 | 0,886 | 0,884 | 0,904 | 0,896 | 0,920 | 0,898 | 0,924 | 0,899 | 0,923 | 0,890 | 0,911 |
| Bayesian Boosting with K-NN | 0,952 | 0,931 | 0,955 | 0,941 | 0,956 | 0,947 | 0,959 | 0,950 | 0,958 | 0,950 | 0,956 | 0,944 |
| AdaBoost with K-NN | 0,952 | 0,931 | 0,956 | 0,941 | 0,957 | 0,948 | 0,958 | 0,950 | 0,957 | 0,949 | 0,956 | 0,944 |
| Bagging with K-NN | 0,961 | 0,948 | 0,966 | 0,957 | 0,966 | 0,963 | 0,967 | 0,964 | 0,968 | 0,962 | 0,966 | 0,959 |

Table C.1.7.: F-measure performance of unbalanced models, for colon cancer.

| F-Measure - Unbalanced Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 93,67% | 91,14% | 95,39% | 94,12% | 96,17% | 95,14% | 97,75% | 97,28% | 97,65% | 97,00% | 96,13% | 94,94% |
| Voting | 93,83% | 91,36% | 95,16% | 93,74% | 95,94% | 95,01% | 97,71% | 97,27% | 97,64% | 96,91% | 96,06% | 94,86% |
| Bayesian Boosting with Decision Tree | 92,93% | 91,24% | 94,57% | 93,37% | 95,57% | 94,78% | 97,41% | 96,89% | 97,18% | 96,64% | 79,61% | 78,82% |
| AdaBoost with Decision Tree | 92,93% | 91,27% | 94,57% | 93,49% | 95,57% | 94,90% | 97,41% | 96,92% | 97,18% | 96,67% | 95,54% | 94,65% |
| Bagging with Decision Tree | 92,40% | 90,43% | 93,88% | 92,35% | 94,71% | 93,78% | 96,72% | 95,94% | 96,74% | 95,55% | 94,89% | 93,61% |
| Bayesian Boosting with Random Forest | 64,83% | 71,30% | 75,06% | 79,82% | 78,12% | 83,69% | 88,94% | 91,62% | 88,33% | 91,42% | 79,06% | 83,57% |
| AdaBoost with Random Forest | 52,77% | 64,09% | 66,23% | 76,57% | 73,62% | 75,46% | 88,26% | 91,23% | 87,64% | 90,69% | 73,70% | 79,61% |
| Bagging with Random Forest | 51,49% | 66,40% | 68,71% | 79,44% | 75,71% | 83,36% | 89,10% | 91,98% | 88,85% | 91,59% | 74,77% | 82,56% |
| Bayesian Boosting with Naive Bayes | 64,93% | 63,60% | 75,73% | 73,87% | 79,69% | 78,77% | 87,61% | 88,30% | 87,23% | 88,30% | 79,04% | 78,57% |
| AdaBoost with Naive Bayes | 67,80% | 62,97% | 75,18% | 74,69% | 78,39% | 78,84% | 87,29% | 88,27% | 87,40% | 88,32% | 79,21% | 78,62% |
| Bagging with Naive Bayes | 65,61% | 62,87% | 73,09% | 74,65% | 74,93% | 78,86% | 82,47% | 88,29% | 81,82% | 88,34% | 75,58% | 78,60% |
| Bayesian Boosting with K-NN | 92,88% | 89,58% | 94,15% | 92,26% | 94,66% | 93,54% | 96,73% | 96,19% | 96,55% | 95,94% | 94,99% | 93,50% |
| AdaBoost with K-NN | 92,88% | 89,58% | 94,15% | 92,26% | 94,66% | 93,54% | 96,73% | 96,19% | 96,55% | 95,94% | 94,99% | 93,50% |
| Bagging with K-NN | 92,88% | 89,45% | 94,12% | 92,09% | 94,64% | 93,62% | 96,74% | 96,21% | 96,66% | 95,95% | 95,01% | 93,46% |

Table C.1.8.: Percentage of wrongly classified cases of unbalanced models, for colon cancer.

| Wrongly Classified Cases - Unbalanced Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 3,06% | 3,63% | 2,69% | 3,68% | 2,85% | 3,19% | 2,64% | 3,19% | 2,80% | 3,76% | 2,81% | 3,49% |
| Voting | 2,69% | 3,71% | 2,82% | 3,52% | 2,98% | 3,37% | 2,57% | 3,47% | 3,29% | 3,76% | 2,87% | 3,57% |
| Bayesian Boosting with Decision Tree | 3,37% | 3,94% | 3,34% | 4,09% | 3,24% | 3,86% | 2,95% | 3,86% | 3,34% | 4,17% | 3,25% | 3,99% |
| AdaBoost with Decision Tree | 3,37% | 3,96% | 3,34% | 4,07% | 3,24% | 3,86% | 2,95% | 3,84% | 3,34% | 4,17% | 3,25% | 3,98% |
| Bagging with Decision Tree | 3,37% | 3,96% | 4,07% | 4,38% | 3,60% | 4,48% | 3,78% | 4,48% | 4,43% | 4,98% | 3,85% | 4,46% |
| Bayesian Boosting with Random Forest | 13,84% | 10,55% | 14,20% | 12,23% | 15,24% | 9,67% | 13,27% | 11,27% | 14,02% | 10,99% | 14,11% | 10,94% |
| AdaBoost with Random Forest | 17,98% | 11,53% | 15,21% | 9,69% | 17,80% | 13,01% | 17,28% | 10,60% | 15,03% | 10,73% | 16,66% | 11,11% |
| Bagging with Random Forest | 17,28% | 12,18% | 14,20% | 10,81% | 13,50% | 11,35% | 13,40% | 10,11% | 14,82% | 9,82% | 14,64% | 10,85% |
| Bayesian Boosting with Naive Bayes | 15,06% | 16,07% | 15,00% | 16,46% | 14,28% | 15,00% | 15,70% | 13,63% | 15,76% | 13,35% | 12,63% | 12,42% |
| AdaBoost with Naive Bayes | 14,38% | 16,46% | 14,93% | 16,51% | 15,13% | 15,11% | 15,37% | 13,63% | 15,88% | 13,35% | 15,14% | 15,01% |
| Bagging with Naive Bayes | 17,05% | 16,35% | 18,48% | 16,51% | 20,06% | 15,03% | 19,85% | 13,71% | 19,98% | 13,35% | 19,08% | 14,99% |
| Bayesian Boosting with K-NN | 3,14% | 4,74% | 3,65% | 4,82% | 3,81% | 4,56% | 4,07% | 4,66% | 4,28% | 4,95% | 3,79% | 4,75% |
| AdaBoost with K-NN | 3,14% | 4,74% | 3,65% | 4,82% | 3,81% | 4,56% | 4,07% | 4,66% | 4,28% | 4,95% | 3,79% | 4,75% |
| Bagging with K-NN | 3,14% | 4,74% | 3,65% | 4,82% | 3,81% | 4,56% | 4,07% | 4,66% | 4,28% | 4,95% | 3,79% | 4,75% |

*Hybrid Models*

Table C.1.9.: Survivability percentage accuracy of hybrid models, for colon cancer.

| Accuracy - Colon Cancer Hybrid Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 98,19% | 96,02% | 98,03% | 96,31% | 98,01% | 96,88% | 97,99% | 97,01% | 97,62% | 96,61% | 97,97% | 96,57% |
| Voting | 97,85% | 95,72% | 97,88% | 96,20% | 97,96% | 96,67% | 98,08% | 96,82% | 97,27% | 96,35% | 97,81% | 96,35% |
| Bayesian Boosting with Decision Tree | 97,59% | 96,21% | 97,44% | 96,38% | 97,62% | 96,83% | 97,83% | 96,97% | 97,45% | 96,64% | 81,32% | 80,51% |
| AdaBoost with Decision Tree | 97,59% | 96,23% | 97,44% | 96,40% | 97,62% | 96,86% | 97,83% | 96,98% | 97,43% | 96,66% | 97,58% | 96,63% |
| Bagging with Decision Tree | 96,52% | 95,13% | 96,69% | 95,58% | 96,76% | 95,79% | 96,95% | 95,89% | 97,06% | 95,34% | 96,80% | 95,55% |
| Bayesian Boosting with Random Forest | 83,47% | 87,15% | 84,34% | 87,83% | 84,79% | 87,92% | 85,20% | 88,85% | 84,66% | 88,93% | 84,49% | 88,14% |
| AdaBoost with Random Forest | 83,30% | 86,01% | 83,14% | 87,37% | 84,20% | 88,51% | 84,23% | 87,92% | 83,33% | 88,64% | 83,64% | 87,69% |
| Bagging with Random Forest | 84,42% | 88,70% | 85,35% | 90,16% | 85,76% | 90,69% | 85,96% | 91,20% | 85,62% | 90,72% | 85,42% | 90,29% |
| Bayesian Boosting with Naive Bayes | 81,71% | 82,09% | 82,62% | 83,59% | 83,33% | 84,18% | 83,60% | 84,87% | 82,96% | 84,80% | 82,84% | 83,91% |
| AdaBoost with Naive Bayes | 81,79% | 81,98% | 82,19% | 83,61% | 84,25% | 84,18% | 83,88% | 84,84% | 83,55% | 84,90% | 83,13% | 83,90% |
| Bagging with Naive Bayes | 80,52% | 81,92% | 79,66% | 83,64% | 80,35% | 84,19% | 79,91% | 84,84% | 80,07% | 84,81% | 80,10% | 83,88% |
| Bayesian Boosting with K-NN | 97,59% | 94,41% | 97,31% | 94,36% | 97,22% | 94,43% | 97,33% | 94,68% | 97,04% | 94,26% | 97,30% | 94,43% |
| AdaBoost with K-NN | 97,58% | 94,38% | 97,26% | 94,29% | 97,27% | 94,45% | 97,29% | 94,67% | 97,07% | 94,26% | 97,29% | 94,41% |
| Bagging with K-NN | 97,58% | 94,38% | 97,35% | 94,47% | 97,27% | 94,49% | 97,26% | 94,69% | 97,08% | 94,17% | 97,31% | 94,44% |

Table C.1.10.: Survivability percentage AUC of hybrid models, for colon cancer.

| AUC - Colon Cancer Hybrid Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 0,990 | 0,983 | 0,992 | 0,985 | 0,993 | 0,989 | 0,994 | 0,991 | 0,993 | 0,990 | 0,992 | 0,988 |
| Voting | 0,987 | 0,978 | 0,988 | 0,980 | 0,987 | 0,983 | 0,989 | 0,985 | 0,988 | 0,984 | 0,988 | 0,982 |
| Bayesian Boosting with Decision Tree | 0,971 | 0,963 | 0,978 | 0,962 | 0,980 | 0,969 | 0,979 | 0,970 | 0,984 | 0,966 | 0,815 | 0,805 |
| AdaBoost with Decision Tree | 0,976 | 0,967 | 0,979 | 0,966 | 0,978 | 0,973 | 0,984 | 0,972 | 0,985 | 0,970 | 0,980 | 0,970 |
| Bagging with Decision Tree | 0,979 | 0,973 | 0,971 | 0,967 | 0,972 | 0,966 | 0,972 | 0,968 | 0,972 | 0,962 | 0,973 | 0,967 |
| Bayesian Boosting with Random Forest | 0,896 | 0,923 | 0,905 | 0,930 | 0,916 | 0,933 | 0,911 | 0,936 | 0,911 | 0,937 | 0,908 | 0,932 |
| AdaBoost with Random Forest | 0,885 | 0,919 | 0,894 | 0,930 | 0,899 | 0,930 | 0,900 | 0,931 | 0,897 | 0,934 | 0,895 | 0,929 |
| Bagging with Random Forest | 0,921 | 0,953 | 0,934 | 0,958 | 0,938 | 0,963 | 0,939 | 0,967 | 0,934 | 0,963 | 0,933 | 0,961 |
| Bayesian Boosting with Naive Bayes | 0,898 | 0,888 | 0,904 | 0,896 | 0,913 | 0,910 | 0,915 | 0,913 | 0,909 | 0,915 | 0,908 | 0,904 |
| AdaBoost with Naive Bayes | 0,899 | 0,889 | 0,907 | 0,899 | 0,916 | 0,910 | 0,916 | 0,915 | 0,914 | 0,915 | 0,910 | 0,906 |
| Bagging with Naive Bayes | 0,872 | 0,886 | 0,884 | 0,902 | 0,896 | 0,921 | 0,899 | 0,925 | 0,897 | 0,922 | 0,890 | 0,911 |
| Bayesian Boosting with K-NN | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 |
| AdaBoost with K-NN | 0,976 | 0,944 | 0,973 | 0,943 | 0,973 | 0,945 | 0,973 | 0,947 | 0,971 | 0,943 | 0,973 | 0,944 |
| Bagging with K-NN | 0,979 | 0,948 | 0,978 | 0,950 | 0,978 | 0,952 | 0,977 | 0,952 | 0,975 | 0,947 | 0,977 | 0,950 |

Table C.1.11.: F-measure performance of hybrid models, for colon cancer.

| F-Measure - Colon Cancer Hybrid Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 98,18% | 95,92% | 98,02% | 96,23% | 98,01% | 96,84% | 97,98% | 96,98% | 97,61% | 96,57% | 97,96% | 96,51% |
| Voting | 97,82% | 95,59% | 97,86% | 96,11% | 97,94% | 96,61% | 98,07% | 96,77% | 97,28% | 96,29% | 97,79% | 96,27% |
| Bayesian Boosting with Decision Tree | 97,56% | 96,15% | 97,41% | 96,34% | 97,60% | 96,79% | 97,82% | 96,94% | 97,43% | 96,61% | 81,30% | 80,47% |
| AdaBoost with Decision Tree | 97,56% | 96,17% | 97,41% | 96,35% | 97,60% | 96,82% | 97,82% | 96,96% | 97,41% | 96,63% | 97,56% | 96,59% |
| Bagging with Decision Tree | 96,42% | 95,00% | 96,60% | 95,48% | 96,68% | 95,70% | 96,89% | 95,80% | 97,00% | 95,23% | 96,72% | 95,44% |
| Bayesian Boosting with Random Forest | 83,71% | 86,87% | 84,70% | 87,86% | 85,46% | 88,12% | 85,57% | 89,05% | 85,12% | 89,18% | 84,91% | 88,22% |
| AdaBoost with Random Forest | 83,41% | 85,83% | 83,64% | 87,35% | 84,87% | 88,86% | 84,89% | 88,18% | 84,11% | 88,95% | 84,19% | 87,83% |
| Bagging with Random Forest | 84,76% | 88,58% | 85,98% | 90,22% | 86,60% | 90,88% | 86,70% | 91,41% | 86,42% | 91,04% | 86,09% | 90,43% |
| Bayesian Boosting with Naive Bayes | 81,92% | 81,89% | 83,00% | 83,57% | 83,95% | 84,43% | 83,97% | 85,12% | 83,57% | 85,18% | 83,28% | 84,04% |
| AdaBoost with Naive Bayes | 82,01% | 81,67% | 82,61% | 83,53% | 85,04% | 84,45% | 84,05% | 85,11% | 84,11% | 85,32% | 83,56% | 84,02% |
| Bagging with Naive Bayes | 80,67% | 81,52% | 79,62% | 83,56% | 79,68% | 84,46% | 79,04% | 85,10% | 79,32% | 85,24% | 79,67% | 83,98% |
| Bayesian Boosting with K-NN | 97,55% | 94,08% | 97,23% | 93,99% | 97,24% | 94,17% | 97,26% | 94,41% | 97,05% | 93,96% | 97,26% | 94,12% |
| AdaBoost with K-NN | 97,55% | 94,08% | 97,23% | 93,99% | 97,24% | 94,17% | 97,26% | 94,41% | 97,05% | 93,96% | 97,26% | 94,12% |
| Bagging with K-NN | 97,55% | 94,08% | 97,32% | 94,19% | 97,24% | 94,21% | 97,24% | 94,43% | 97,05% | 93,85% | 97,28% | 94,15% |

Table C.1.12.: Percentage of wrongly classified cases of hybrid models, for colon cancer.

| Wrongly Classified Cases - Colon Cancer Hybrid Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 13,97% | 41,02% | 12,65% | 60,33% | 21,02% | 58,95% | 10,88% | 49,49% | 16,30% | 48,30% | 14,96% | 51,62% |
| Voting | 22,83% | 47,53% | 35,27% | 60,12% | 47,45% | 58,43% | 44,08% | 53,36% | 52,81% | 53,56% | 40,49% | 54,60% |
| Bayesian Boosting with Decision Tree | 71,94% | 43,85% | 65,56% | 53,82% | 57,61% | 58,10% | 66,00% | 50,66% | 58,98% | 46,20% | 64,02% | 50,53% |
| AdaBoost with Decision Tree | 71,94% | 43,98% | 65,56% | 53,80% | 57,61% | 58,10% | 66,00% | 51,31% | 58,98% | 44,96% | 64,02% | 50,43% |
| Bagging with Decision Tree | 74,37% | 51,46% | 66,21% | 59,86% | 61,57% | 57,94% | 63,80% | 54,78% | 59,60% | 43,92% | 65,11% | 53,59% |
| Bayesian Boosting with Random Forest | 37,34% | 54,39% | 56,96% | 64,27% | 58,98% | 65,51% | 47,97% | 47,73% | 56,85% | 56,05% | 51,62% | 57,59% |
| AdaBoost with Random Forest | 65,07% | 60,77% | 66,36% | 62,24% | 64,50% | 55,35% | 49,31% | 47,47% | 59,55% | 59,55% | 60,96% | 57,08% |
| Bagging with Random Forest | 43,43% | 51,98% | 61,41% | 64,76% | 61,60% | 58,43% | 31,54% | 46,98% | 60,09% | 50,76% | 51,61% | 54,58% |
| Bayesian Boosting with Naive Bayes | 42,65% | 42,86% | 63,67% | 68,26% | 57,22% | 62,17% | 61,47% | 49,70% | 53,56% | 48,95% | 46,43% | 45,32% |
| AdaBoost with Naive Bayes | 42,89% | 42,50% | 63,80% | 68,07% | 57,45% | 62,17% | 61,47% | 49,70% | 53,59% | 49,42% | 55,84% | 54,37% |
| Bagging with Naive Bayes | 42,37% | 42,99% | 58,38% | 68,26% | 57,58% | 62,14% | 56,96% | 49,60% | 51,18% | 48,72% | 53,29% | 54,34% |
| Bayesian Boosting with K-NN | 12,13% | 55,43% | 11,09% | 62,48% | 11,95% | 59,21% | 10,13% | 54,21% | 13,63% | 55,79% | 11,79% | 57,42% |
| AdaBoost with K-NN | 12,13% | 55,43% | 11,09% | 62,48% | 11,95% | 59,21% | 10,13% | 54,21% | 13,63% | 55,79% | 11,79% | 57,42% |
| Bagging with K-NN | 12,13% | 55,43% | 11,09% | 62,48% | 11,95% | 59,21% | 10,13% | 54,21% | 13,63% | 55,79% | 11,79% | 57,42% |

*Oversampled Models*

Table C.1.13.: Survivability percentage accuracy of oversampled models, for colon cancer.

| Accuracy - Colon Cancer Oversampled Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 98,84% | 96,54% | 97,77% | 97,17% | 98,33% | 97,24% | 98,28% | 97,37% | 97,91% | 96,90% | 98,23% | 97,04% |
| Voting | 98,58% | 96,45% | 97,61% | 96,87% | 98,39% | 96,93% | 98,30% | 97,19% | 98,14% | 96,66% | 98,20% | 96,82% |
| Bayesian Boosting with Decision Tree | 98,48% | 96,96% | 97,37% | 97,14% | 98,11% | 97,20% | 98,11% | 97,37% | 97,84% | 96,95% | 81,65% | 80,94% |
| AdaBoost with Decision Tree | 98,48% | 96,97% | 97,37% | 97,16% | 98,11% | 97,24% | 98,11% | 97,39% | 97,84% | 96,97% | 97,98% | 97,15% |
| Bagging with Decision Tree | 97,68% | 96,25% | 97,06% | 96,30% | 97,53% | 96,42% | 97,37% | 96,32% | 97,16% | 95,79% | 97,36% | 96,22% |
| Bayesian Boosting with Random Forest | 83,21% | 86,96% | 84,28% | 87,99% | 85,05% | 88,42% | 84,85% | 87,93% | 84,68% | 88,65% | 84,41% | 87,99% |
| AdaBoost with Random Forest | 82,93% | 86,83% | 84,19% | 88,13% | 84,15% | 87,29% | 83,81% | 89,13% | 83,62% | 88,27% | 83,74% | 87,93% |
| Bagging with Random Forest | 85,09% | 88,74% | 85,40% | 90,12% | 85,87% | 90,87% | 85,90% | 91,12% | 86,02% | 90,77% | 85,66% | 90,32% |
| Bayesian Boosting with Naive Bayes | 81,89% | 82,31% | 83,08% | 83,63% | 83,31% | 84,58% | 83,15% | 84,71% | 82,87% | 84,91% | 82,86% | 84,03% |
| AdaBoost with Naive Bayes | 82,12% | 82,15% | 82,78% | 83,62% | 83,36% | 84,58% | 84,16% | 84,64% | 83,58% | 84,97% | 83,20% | 83,99% |
| Bagging with Naive Bayes | 80,76% | 82,23% | 79,63% | 83,59% | 80,31% | 84,57% | 79,95% | 84,59% | 80,16% | 84,90% | 80,16% | 83,98% |
| Bayesian Boosting with K-NN | 98,45% | 95,19% | 98,10% | 95,30% | 97,67% | 94,92% | 97,58% | 95,09% | 97,45% | 94,44% | 97,85% | 94,99% |
| AdaBoost with K-NN | 98,45% | 95,19% | 98,10% | 95,30% | 97,67% | 94,92% | 97,58% | 95,09% | 97,45% | 94,44% | 97,85% | 94,99% |
| Bagging with K-NN | 98,45% | 95,12% | 98,11% | 95,25% | 97,68% | 94,97% | 97,56% | 95,05% | 97,41% | 94,45% | 97,84% | 94,97% |

Table C.1.14.: Survivability percentage AUC of oversampled models, for colon cancer.

| AUC - Colon Cancer Oversampled Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 0,982 | 0,986 | 0,993 | 0,988 | 0,995 | 0,991 | 0,994 | 0,992 | 0,994 | 0,991 | 0,992 | 0,990 |
| Voting | 0,992 | 0,982 | 0,989 | 0,983 | 0,991 | 0,985 | 0,990 | 0,986 | 0,990 | 0,985 | 0,990 | 0,984 |
| Bayesian Boosting with Decision Tree | 0,982 | 0,969 | 0,991 | 0,974 | 0,982 | 0,970 | 0,978 | 0,971 | 0,982 | 0,973 | 0,819 | 0,810 |
| AdaBoost with Decision Tree | 0,986 | 0,971 | 0,990 | 0,974 | 0,986 | 0,975 | 0,984 | 0,976 | 0,985 | 0,974 | 0,986 | 0,974 |
| Bagging with Decision Tree | 0,989 | 0,984 | 0,982 | 0,977 | 0,980 | 0,975 | 0,981 | 0,976 | 0,978 | 0,970 | 0,982 | 0,976 |
| Bayesian Boosting with Random Forest | 0,899 | 0,928 | 0,908 | 0,935 | 0,912 | 0,940 | 0,911 | 0,936 | 0,905 | 0,938 | 0,907 | 0,935 |
| AdaBoost with Random Forest | 0,883 | 0,919 | 0,903 | 0,931 | 0,898 | 0,925 | 0,896 | 0,935 | 0,902 | 0,931 | 0,896 | 0,928 |
| Bagging with Random Forest | 0,926 | 0,954 | 0,934 | 0,959 | 0,938 | 0,963 | 0,938 | 0,966 | 0,939 | 0,964 | 0,935 | 0,961 |
| Bayesian Boosting with Naive Bayes | 0,897 | 0,888 | 0,903 | 0,897 | 0,915 | 0,912 | 0,912 | 0,913 | 0,910 | 0,915 | 0,907 | 0,905 |
| AdaBoost with Naive Bayes | 0,900 | 0,889 | 0,905 | 0,899 | 0,915 | 0,911 | 0,916 | 0,915 | 0,915 | 0,915 | 0,910 | 0,906 |
| Bagging with Naive Bayes | 0,872 | 0,886 | 0,883 | 0,903 | 0,897 | 0,920 | 0,898 | 0,925 | 0,899 | 0,924 | 0,890 | 0,912 |
| Bayesian Boosting with K-NN | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 |
| AdaBoost with K-NN | 0,984 | 0,952 | 0,981 | 0,953 | 0,977 | 0,949 | 0,976 | 0,951 | 0,975 | 0,944 | 0,979 | 0,950 |
| Bagging with K-NN | 0,986 | 0,953 | 0,983 | 0,955 | 0,980 | 0,955 | 0,979 | 0,957 | 0,978 | 0,951 | 0,981 | 0,954 |

Table C.1.15.: F-measure performance of oversampled models, for colon cancer.

| F-Measure - Colon Cancer Oversampled Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 98,83% | 96,45% | 97,77% | 97,13% | 98,32% | 97,21% | 98,27% | 97,34% | 97,91% | 96,87% | 98,22% | 97,00% |
| Voting | 98,56% | 96,36% | 97,61% | 96,81% | 98,37% | 96,88% | 98,29% | 97,15% | 98,13% | 96,61% | 98,19% | 96,76% |
| Bayesian Boosting with Decision Tree | 98,46% | 96,93% | 97,37% | 97,11% | 98,10% | 97,18% | 98,10% | 98,14% | 97,83% | 96,92% | 81,64% | 81,05% |
| AdaBoost with Decision Tree | 98,46% | 96,94% | 97,37% | 97,13% | 98,10% | 97,21% | 98,10% | 97,37% | 97,83% | 96,94% | 97,97% | 97,12% |
| Bagging with Decision Tree | 97,62% | 96,18% | 97,02% | 96,22% | 97,49% | 96,35% | 97,33% | 96,25% | 97,10% | 95,69% | 97,31% | 96,14% |
| Bayesian Boosting with Random Forest | 83,39% | 86,75% | 84,73% | 88,03% | 85,43% | 88,52% | 85,27% | 88,13% | 85,05% | 88,88% | 84,78% | 88,06% |
| AdaBoost with Random Forest | 83,27% | 86,59% | 84,68% | 88,20% | 84,68% | 87,68% | 84,44% | 89,28% | 84,31% | 88,63% | 84,28% | 88,08% |
| Bagging with Random Forest | 85,42% | 88,60% | 86,11% | 90,25% | 86,61% | 91,05% | 86,58% | 91,35% | 86,83% | 91,07% | 86,31% | 90,46% |
| Bayesian Boosting with Naive Bayes | 82,05% | 82,01% | 83,47% | 83,60% | 83,77% | 84,76% | 83,26% | 84,98% | 83,42% | 85,27% | 83,19% | 84,12% |
| AdaBoost with Naive Bayes | 82,37% | 81,80% | 83,11% | 83,55% | 83,71% | 84,77% | 84,70% | 84,94% | 84,25% | 85,39% | 83,63% | 84,09% |
| Bagging with Naive Bayes | 80,85% | 81,77% | 79,57% | 83,52% | 79,62% | 84,77% | 79,06% | 84,88% | 79,35% | 85,29% | 79,69% | 84,04% |
| Bayesian Boosting with K-NN | 98,43% | 94,96% | 98,08% | 95,09% | 97,64% | 94,67% | 97,56% | 94,87% | 97,43% | 94,15% | 97,83% | 94,75% |
| AdaBoost with K-NN | 98,43% | 94,96% | 98,08% | 95,09% | 97,64% | 94,67% | 97,56% | 94,87% | 97,43% | 94,15% | 97,83% | 94,75% |
| Bagging with K-NN | 98,43% | 94,88% | 98,10% | 95,03% | 97,65% | 94,73% | 97,54% | 94,83% | 97,39% | 94,16% | 97,82% | 94,73% |

Table C.1.16.: Percentage of wrongly classified cases of oversampled models, for colon cancer.

| Wrongly Classified Cases - Colon Cancer Oversampled Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 13,35% | 47,53% | 12,65% | 59,34% | 24,51% | 56,47% | 23,76% | 59,26% | 22,62% | 61,31% | 19,38% | 56,78% |
| Voting | 42,34% | 51,65% | 47,73% | 61,41% | 39,70% | 57,40% | 59,32% | 61,93% | 55,61% | 60,46% | 48,94% | 58,57% |
| Bayesian Boosting with Decision Tree | 72,04% | 45,76% | 64,14% | 54,94% | 62,63% | 54,81% | 58,69% | 58,43% | 58,41% | 59,94% | 63,18% | 54,78% |
| AdaBoost with Decision Tree | 72,04% | 45,79% | 64,14% | 54,91% | 62,63% | 54,73% | 58,69% | 58,64% | 58,41% | 59,99% | 63,18% | 54,81% |
| Bagging with Decision Tree | 74,94% | 50,82% | 66,65% | 57,86% | 63,25% | 57,22% | 61,70% | 58,15% | 59,47% | 57,42% | 65,20% | 56,29% |
| Bayesian Boosting with Random Forest | 56,49% | 53,67% | 52,79% | 65,07% | 51,28% | 55,77% | 55,20% | 60,90% | 58,72% | 65,72% | 54,90% | 60,22% |
| AdaBoost with Random Forest | 40,40% | 52,11% | 64,71% | 62,14% | 43,48% | 62,22% | 42,16% | 60,20% | 59,03% | 66,23% | 49,96% | 60,58% |
| Bagging with Random Forest | 51,80% | 53,02% | 64,32% | 65,79% | 64,06% | 58,46% | 59,26% | 62,50% | 59,70% | 62,11% | 59,83% | 60,38% |
| Bayesian Boosting with Naive Bayes | 40,79% | 52,29% | 59,83% | 68,02% | 62,30% | 62,89% | 53,80% | 64,63% | 53,36% | 65,87% | 45,01% | 52,28% |
| AdaBoost with Naive Bayes | 39,52% | 51,93% | 60,56% | 68,10% | 62,32% | 62,89% | 53,80% | 64,63% | 53,28% | 66,31% | 53,89% | 62,77% |
| Bagging with Naive Bayes | 40,53% | 52,32% | 56,93% | 68,00% | 63,51% | 62,53% | 54,68% | 64,55% | 53,10% | 65,77% | 53,75% | 62,63% |
| Bayesian Boosting with K-NN | 11,64% | 55,90% | 11,92% | 62,04% | 12,80% | 60,17% | 13,24% | 60,12% | 12,88% | 61,41% | 12,50% | 59,93% |
| AdaBoost with K-NN | 11,64% | 55,90% | 11,92% | 62,04% | 12,80% | 60,17% | 13,24% | 60,12% | 12,88% | 61,41% | 12,50% | 59,93% |
| Bagging with K-NN | 11,64% | 55,90% | 11,92% | 62,04% | 12,80% | 60,17% | 13,24% | 60,12% | 12,88% | 61,41% | 12,50% | 59,93% |

*Undersampled Models*

Table C.1.17.: Survivability percentage accuracy of undersampled models, for colon cancer.

| Accuracy - Colon Cancer Undersampled Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 94,45% | 93,96% | 95,56% | 95,00% | 96,06% | 95,68% | 96,36% | 96,12% | 97,94% | 96,99% | 96,07% | 95,55% |
| Voting | 95,29% | 94,10% | 96,03% | 94,89% | 96,55% | 95,61% | 96,79% | 96,13% | 96,65% | 95,70% | 96,26% | 95,29% |
| Bayesian Boosting with Decision Tree | 94,58% | 93,78% | 95,34% | 94,72% | 95,94% | 95,46% | 96,35% | 95,95% | 96,28% | 95,69% | 79,75% | 79,27% |
| AdaBoost with Decision Tree | 94,58% | 93,82% | 95,34% | 94,76% | 95,94% | 95,52% | 96,35% | 95,98% | 96,28% | 95,76% | 95,70% | 95,17% |
| Bagging with Decision Tree | 93,67% | 92,87% | 94,58% | 93,77% | 95,24% | 94,57% | 95,72% | 95,32% | 95,76% | 94,77% | 94,99% | 94,26% |
| Bayesian Boosting with Random Forest | 82,37% | 86,22% | 83,54% | 86,90% | 84,43% | 87,53% | 85,17% | 88,70% | 84,54% | 88,38% | 84,01% | 87,55% |
| AdaBoost with Random Forest | 83,13% | 85,03% | 82,41% | 86,83% | 83,63% | 87,45% | 84,29% | 87,21% | 82,79% | 87,21% | 83,25% | 86,75% |
| Bagging with Random Forest | 84,32% | 87,63% | 84,27% | 89,60% | 85,31% | 90,34% | 85,70% | 90,64% | 85,18% | 90,24% | 84,96% | 89,69% |
| Bayesian Boosting with Naive Bayes | 81,59% | 82,19% | 82,85% | 83,41% | 83,03% | 84,35% | 84,10% | 84,78% | 82,41% | 84,89% | 82,80% | 83,92% |
| AdaBoost with Naive Bayes | 81,20% | 82,19% | 82,89% | 83,37% | 83,38% | 84,36% | 83,78% | 84,73% | 83,48% | 85,07% | 82,95% | 83,94% |
| Bagging with Naive Bayes | 80,62% | 82,26% | 79,79% | 83,41% | 80,37% | 84,36% | 79,89% | 84,69% | 80,19% | 84,98% | 80,17% | 83,94% |
| Bayesian Boosting with K-NN | 94,17% | 91,23% | 94,63% | 92,46% | 95,20% | 93,74% | 95,36% | 94,24% | 95,35% | 94,07% | 94,94% | 93,15% |
| AdaBoost with K-NN | 94,17% | 91,23% | 94,63% | 92,46% | 95,20% | 93,74% | 95,36% | 94,24% | 95,35% | 94,07% | 94,94% | 93,15% |
| Bagging with K-NN | 94,17% | 91,00% | 94,69% | 92,49% | 95,12% | 93,90% | 95,29% | 94,34% | 95,29% | 93,99% | 94,91% | 93,14% |

Table C.1.18.: Survivability percentage AUC of undersampled models, for colon cancer.

| AUC - Colon Cancer Undersampled Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 0,980 | 0,978 | 0,985 | 0,983 | 0,988 | 0,985 | 0,989 | 0,986 | 0,994 | 0,992 | 0,987 | 0,985 |
| Voting | 0,969 | 0,963 | 0,976 | 0,970 | 0,980 | 0,974 | 0,981 | 0,977 | 0,981 | 0,975 | 0,977 | 0,972 |
| Bayesian Boosting with Decision Tree | 0,938 | 0,940 | 0,955 | 0,941 | 0,960 | 0,961 | 0,959 | 0,961 | 0,975 | 0,958 | 0,798 | 0,794 |
| AdaBoost with Decision Tree | 0,957 | 0,943 | 0,966 | 0,951 | 0,969 | 0,958 | 0,965 | 0,962 | 0,977 | 0,959 | 0,967 | 0,955 |
| Bagging with Decision Tree | 0,909 | 0,900 | 0,936 | 0,916 | 0,944 | 0,928 | 0,949 | 0,927 | 0,948 | 0,924 | 0,937 | 0,919 |
| Bayesian Boosting with Random Forest | 0,888 | 0,922 | 0,897 | 0,924 | 0,905 | 0,933 | 0,910 | 0,937 | 0,905 | 0,929 | 0,901 | 0,929 |
| AdaBoost with Random Forest | 0,890 | 0,911 | 0,894 | 0,923 | 0,901 | 0,931 | 0,899 | 0,928 | 0,892 | 0,927 | 0,895 | 0,924 |
| Bagging with Random Forest | 0,919 | 0,946 | 0,928 | 0,956 | 0,935 | 0,961 | 0,936 | 0,964 | 0,935 | 0,960 | 0,931 | 0,957 |
| Bayesian Boosting with Naive Bayes | 0,897 | 0,887 | 0,902 | 0,896 | 0,914 | 0,909 | 0,916 | 0,912 | 0,908 | 0,913 | 0,907 | 0,903 |
| AdaBoost with Naive Bayes | 0,897 | 0,887 | 0,904 | 0,898 | 0,915 | 0,910 | 0,914 | 0,913 | 0,913 | 0,915 | 0,909 | 0,905 |
| Bagging with Naive Bayes | 0,870 | 0,887 | 0,883 | 0,902 | 0,897 | 0,919 | 0,899 | 0,924 | 0,900 | 0,924 | 0,890 | 0,911 |
| Bayesian Boosting with K-NN | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 |
| AdaBoost with K-NN | 0,942 | 0,912 | 0,946 | 0,925 | 0,952 | 0,937 | 0,954 | 0,942 | 0,954 | 0,941 | 0,950 | 0,931 |
| Bagging with K-NN | 0,955 | 0,930 | 0,958 | 0,943 | 0,962 | 0,954 | 0,964 | 0,957 | 0,963 | 0,953 | 0,960 | 0,947 |

Table C.1.19.: F-measure performance of undersampled models, for colon cancer.

| F-Measure - Colon Cancer Undersampled Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 94,55% | 93,89% | 95,62% | 94,95% | 96,11% | 95,64% | 96,41% | 96,08% | 97,94% | 96,97% | 96,13% | 95,51% |
| Voting | 95,27% | 93,98% | 96,01% | 94,78% | 96,53% | 95,52% | 96,77% | 96,05% | 96,63% | 95,59% | 96,24% | 95,18% |
| Bayesian Boosting with Decision Tree | 94,55% | 93,72% | 95,32% | 94,67% | 95,92% | 95,41% | 96,32% | 95,90% | 96,26% | 95,63% | 79,73% | 79,22% |
| AdaBoost with Decision Tree | 94,55% | 93,76% | 95,32% | 94,72% | 95,92% | 95,47% | 96,32% | 95,94% | 96,26% | 95,70% | 95,68% | 95,12% |
| Bagging with Decision Tree | 93,45% | 92,64% | 94,41% | 93,60% | 95,12% | 94,40% | 95,63% | 95,21% | 95,66% | 94,59% | 94,85% | 94,09% |
| Bayesian Boosting with Random Forest | 81,89% | 86,09% | 82,92% | 86,66% | 83,65% | 87,22% | 84,57% | 88,30% | 84,13% | 87,85% | 83,43% | 87,23% |
| AdaBoost with Random Forest | 82,61% | 84,93% | 81,30% | 86,27% | 82,31% | 86,84% | 83,38% | 86,64% | 81,29% | 86,39% | 82,18% | 86,22% |
| Bagging with Random Forest | 83,97% | 87,57% | 83,42% | 89,41% | 84,35% | 90,02% | 84,81% | 90,29% | 84,15% | 89,78% | 84,14% | 89,42% |
| Bayesian Boosting with Naive Bayes | 81,41% | 82,51% | 82,55% | 83,48% | 82,33% | 84,16% | 83,51% | 84,54% | 81,88% | 84,49% | 82,34% | 83,84% |
| AdaBoost with Naive Bayes | 80,99% | 82,63% | 82,63% | 83,49% | 82,93% | 84,16% | 83,64% | 84,47% | 82,98% | 84,61% | 82,63% | 83,87% |
| Bagging with Naive Bayes | 80,56% | 82,74% | 79,84% | 83,52% | 81,06% | 84,16% | 80,74% | 84,41% | 80,88% | 84,52% | 80,62% | 83,87% |
| Bayesian Boosting with K-NN | 94,13% | 90,63% | 94,60% | 92,01% | 95,16% | 93,45% | 95,33% | 93,99% | 95,31% | 93,81% | 94,91% | 92,78% |
| AdaBoost with K-NN | 94,13% | 90,63% | 94,60% | 92,01% | 95,16% | 93,45% | 95,33% | 93,99% | 95,31% | 93,81% | 94,91% | 92,78% |
| Bagging with K-NN | 94,13% | 90,37% | 94,67% | 92,04% | 95,09% | 93,63% | 95,26% | 94,09% | 95,25% | 93,72% | 94,88% | 92,77% |

Table C.1.20.: Percentage of wrongly classified cases of undersampled models, for colon cancer.

| Wrongly Classified Cases - Colon Cancer Undersampled Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 26,79% | 32,05% | 23,79% | 51,88% | 62,79% | 53,90% | 59,96% | 47,47% | 22,86% | 62,45% | 39,24% | 49,55% |
| Voting | 13,29% | 27,68% | 16,71% | 48,56% | 35,35% | 50,01% | 38,07% | 46,41% | 37,16% | 49,16% | 28,12% | 44,36% |
| Bayesian Boosting with Decision Tree | 63,95% | 33,58% | 56,47% | 44,42% | 65,61% | 48,41% | 63,13% | 45,32% | 60,38% | 47,11% | 61,91% | 43,77% |
| AdaBoost with Decision Tree | 63,95% | 33,66% | 56,47% | 44,31% | 65,61% | 48,92% | 63,13% | 45,37% | 60,38% | 47,91% | 61,91% | 44,04% |
| Bagging with Decision Tree | 59,89% | 32,24% | 55,09% | 38,87% | 67,82% | 48,67% | 45,69% | 62,43% | 50,56% | 61,03% | 55,81% | 48,64% |
| Bayesian Boosting with Random Forest | 28,38% | 31,30% | 31,54% | 59,96% | 49,83% | 50,66% | 34,62% | 41,77% | 44,60% | 50,01% | 37,79% | 46,74% |
| AdaBoost with Random Forest | 30,22% | 29,15% | 35,03% | 55,90% | 30,99% | 52,89% | 30,40% | 34,96% | 46,33% | 46,13% | 34,59% | 43,80% |
| Bagging with Random Forest | 17,39% | 28,12% | 30,60% | 52,22% | 29,75% | 52,35% | 28,58% | 47,34% | 34,08% | 48,74% | 28,08% | 45,75% |
| Bayesian Boosting with Naive Bayes | 52,19% | 33,64% | 48,77% | 65,74% | 63,95% | 62,30% | 61,00% | 48,59% | 54,39% | 51,41% | 46,72% | 43,61% |
| AdaBoost with Naive Bayes | 52,09% | 33,58% | 48,69% | 63,70% | 63,82% | 62,30% | 61,03% | 48,59% | 54,57% | 50,76% | 56,04% | 51,79% |
| Bagging with Naive Bayes | 35,86% | 33,53% | 42,83% | 65,64% | 62,92% | 62,30% | 59,83% | 48,59% | 50,32% | 51,59% | 50,36% | 52,33% |
| Bayesian Boosting with K-NN | 8,01% | 27,39% | 8,68% | 42,78% | 7,93% | 45,22% | 9,17% | 43,41% | 9,67% | 47,73% | 8,69% | 41,31% |
| AdaBoost with K-NN | 8,01% | 27,39% | 8,68% | 42,78% | 7,93% | 45,22% | 9,17% | 43,41% | 9,67% | 47,73% | 8,69% | 41,31% |
| Bagging with K-NN | 8,01% | 27,39% | 8,68% | 42,78% | 7,93% | 45,22% | 9,17% | 43,41% | 9,67% | 47,73% | 8,69% | 41,31% |

### C.1.2 *Rectal Cancer*

*Unbalanced Models Without Unknowns*

Table C.1.21.: Survivability percentage accuracy of unbalanced models without unknowns, for rectal cancer.

| Accuracy - Rectal Cancer Unbalanced Models Without Unknowns | | | | | | |
|---|---|---|---|---|---|---|
| | 1Year | 2Year | 3Year | 4Year | 5Year | Average |
| | 6 attributes | 6 attributes | 6 attributes | 6 attributes | 6 attributes | 6 attributes |
| Stacking | 94,36% | 94,94% | 94,70% | 94,46% | 93,77% | 94,45% |
| Voting | 96,07% | 95,10% | 94,80% | 94,20% | 93,83% | 94,80% |
| Bayesian Boosting with Decision Tree | 95,34% | 94,64% | 94,14% | 92,91% | 92,66% | 78,28% |
| AdaBoost with Decision Tree | 95,34% | 94,64% | 94,14% | 92,91% | 92,66% | 93,94% |
| Bagging with Decision Tree | 95,14% | 93,81% | 92,83% | 90,47% | 90,15% | 92,48% |
| Bayesian Boosting with Random Forest | 91,74% | 90,04% | 89,10% | 88,39% | 87,28% | 89,31% |
| AdaBoost with Random Forest | 90,89% | 88,57% | 87,77% | 87,53% | 85,17% | 87,99% |
| Bagging with Random Forest | 90,65% | 89,70% | 88,25% | 87,71% | 86,56% | 88,57% |
| Bayesian Boosting with Naive Bayes | 89,96% | 88,78% | 86,54% | 85,67% | 83,94% | 86,98% |
| AdaBoost with Naive Bayes | 89,46% | 88,29% | 86,74% | 85,31% | 83,66% | 86,69% |
| Bagging with Naive Bayes | 89,08% | 87,87% | 85,95% | 85,49% | 83,90% | 86,46% |
| Bayesian Boosting with K-NN | 95,10% | 93,91% | 93,37% | 92,54% | 92,10% | 93,40% |
| AdaBoost with K-NN | 95,22% | 94,08% | 93,29% | 92,54% | 91,92% | 93,41% |
| Bagging with K-NN | 94,74% | 93,85% | 93,11% | 92,77% | 92,12% | 93,32% |

Table C.1.22.: Survivability percentage AUC of unbalanced models without unknowns, for rectal cancer.

| AUC - Rectal Cancer Unbalanced Models Without Unknowns | | | | | | |
|---|---|---|---|---|---|---|
| | 1Year | 2Year | 3Year | 4Year | 5Year | Average |
| | 6 attributes | 6 attributes | 6 attributes | 6 attributes | 6 attributes | 6 attributes |
| Stacking | 0,957 | 0,968 | 0,968 | 0,968 | 0,963 | 0,965 |
| Voting | 0,896 | 0,927 | 0,939 | 0,900 | 0,914 | 0,915 |
| Bayesian Boosting with Decision Tree | 0,854 | 0,899 | 0,888 | 0,907 | 0,914 | 0,744 |
| AdaBoost with Decision Tree | 0,873 | 0,891 | 0,912 | 0,920 | 0,923 | 0,904 |
| Bagging with Decision Tree | 0,883 | 0,896 | 0,891 | 0,762 | 0,800 | 0,846 |
| Bayesian Boosting with Random Forest | 0,708 | 0,797 | 0,819 | 0,831 | 0,826 | 0,796 |
| AdaBoost with Random Forest | 0,652 | 0,760 | 0,791 | 0,817 | 0,798 | 0,764 |
| Bagging with Random Forest | 0,942 | 0,933 | 0,935 | 0,928 | 0,920 | 0,932 |
| Bayesian Boosting with Naive Bayes | 0,868 | 0,884 | 0,871 | 0,867 | 0,864 | 0,871 |
| AdaBoost with Naive Bayes | 0,856 | 0,871 | 0,873 | 0,868 | 0,864 | 0,866 |
| Bagging with Naive Bayes | 0,876 | 0,892 | 0,880 | 0,870 | 0,860 | 0,876 |
| Bayesian Boosting with K-NN | 0,500 | 0,877 | 0,887 | 0,885 | 0,886 | 0,807 |
| AdaBoost with K-NN | 0,869 | 0,878 | 0,887 | 0,886 | 0,884 | 0,881 |
| Bagging with K-NN | 0,871 | 0,896 | 0,906 | 0,905 | 0,903 | 0,896 |

Table C.1.23.: F-measure performance of unbalanced models without unknowns, for rectal cancer.

| F-Measure - Rectal Cancer Unbalanced Models Without Unknowns | | | | | | |
|---|---|---|---|---|---|---|
| | 1Year | 2Year | 3Year | 4Year | 5Year | Average |
| | 6 attributes | 6 attributes | 6 attributes | 6 attributes | 6 attributes | 6 attributes |
| Stacking | 76,22% | 84,79% | 87,08% | 96,41% | 95,87% | 88,08% |
| Voting | 80,02% | 83,90% | 86,54% | 96,31% | 95,98% | 88,55% |
| Bayesian Boosting with Decision Tree | 77,91% | 82,99% | 85,27% | 95,37% | 95,09% | 72,77% |
| AdaBoost with Decision Tree | 77,91% | 82,99% | 85,27% | 95,37% | 95,09% | 87,33% |
| Bagging with Decision Tree | 74,69% | 79,04% | 80,56% | 93,59% | 93,21% | 84,22% |
| Bayesian Boosting with Random Forest | 45,44% | 61,74% | 66,63% | 92,85% | 92,01% | 71,73% |
| AdaBoost with Random Forest | 34,23% | 53,38% | 61,36% | 92,34% | 90,83% | 66,43% |
| Bagging with Random Forest | 28,43% | 58,41% | 62,00% | 92,53% | 91,69% | 66,61% |
| Bayesian Boosting with Naive Bayes | 52,15% | 63,89% | 63,29% | 91,00% | 89,72% | 72,01% |
| AdaBoost with Naive Bayes | 48,43% | 63,26% | 64,96% | 90,69% | 89,43% | 71,35% |
| Bagging with Naive Bayes | 54,51% | 64,01% | 64,54% | 90,78% | 89,56% | 72,68% |
| Bayesian Boosting with K-NN | 77,85% | 81,52% | 83,23% | 95,21% | 94,68% | 86,50% |
| AdaBoost with K-NN | 77,85% | 81,52% | 83,23% | 95,21% | 94,68% | 86,50% |
| Bagging with K-NN | 75,36% | 81,02% | 82,75% | 95,35% | 94,81% | 85,86% |

Table C.1.24.: Percentage of wrongly classified cases of unbalanced models without unknowns, for rectal cancer.

| Wrongly Classified Cases - Rectal Cancer Unbalanced Models Without Unknowns | | | | | | |
|---|---|---|---|---|---|---|
| | 1Year | 2Year | 3Year | 4Year | 5Year | Average |
| | 6 attributes | 6 attributes | 6 attributes | 6 attributes | 6 attributes | 6 attributes |
| Stacking | 5,08% | 4,54% | 4,17% | 5,08% | 5,08% | 4,79% |
| Voting | 3,99% | 4,54% | 4,36% | 5,81% | 5,81% | 4,90% |
| Bayesian Boosting with Decision Tree | 4,54% | 5,81% | 5,44% | 6,35% | 6,90% | 5,81% |
| AdaBoost with Decision Tree | 4,54% | 5,81% | 5,44% | 6,35% | 6,90% | 5,81% |
| Bagging with Decision Tree | 4,72% | 5,63% | 6,53% | 9,98% | 9,26% | 7,22% |
| Bayesian Boosting with Random Forest | 7,44% | 9,26% | 10,34% | 12,34% | 12,70% | 10,42% |
| AdaBoost with Random Forest | 8,53% | 10,16% | 12,16% | 11,07% | 13,07% | 11,00% |
| Bagging with Random Forest | 7,99% | 9,07% | 9,80% | 11,98% | 13,07% | 10,38% |
| Bayesian Boosting with Naive Bayes | 8,53% | 10,89% | 11,25% | 12,89% | 15,43% | 9,83% |
| AdaBoost with Naive Bayes | 8,71% | 10,89% | 12,16% | 14,88% | 15,97% | 12,52% |
| Bagging with Naive Bayes | 9,98% | 12,70% | 12,34% | 14,16% | 15,97% | 13,03% |
| Bayesian Boosting with K-NN | 4,72% | 5,63% | 5,08% | 6,72% | 6,72% | 5,77% |
| AdaBoost with K-NN | 4,72% | 5,63% | 5,08% | 6,72% | 6,72% | 5,77% |
| Bagging with K-NN | 4,72% | 5,63% | 5,08% | 6,72% | 6,72% | 5,77% |

*Unbalanced Models*

Table C.1.25.: Survivability percentage accuracy of unbalanced models, for rectal cancer.

| Accuracy - Rectal Cancer Unbalanced Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 94,34% | 94,42% | 94,14% | 94,45% | 93,66% | 94,05% | 94,51% | 93,89% | 94,69% | 94,51% | 94,27% | 94,13% |
| Voting | 97,01% | 95,20% | 96,98% | 95,27% | 97,07% | 95,42% | 97,16% | 95,68% | 97,20% | 95,38% | 97,08% | 95,39% |
| Bayesian Boosting with Decision Tree | 96,55% | 95,24% | 96,48% | 95,24% | 96,74% | 95,27% | 96,87% | 95,52% | 96,69% | 95,57% | 80,56% | 79,47% |
| AdaBoost with Decision Tree | 96,55% | 95,24% | 96,48% | 95,24% | 96,74% | 95,27% | 96,87% | 95,52% | 96,69% | 95,57% | 96,67% | 95,37% |
| Bagging with Decision Tree | 96,42% | 95,10% | 96,14% | 94,71% | 96,20% | 94,76% | 96,09% | 94,91% | 96,21% | 95,13% | 96,21% | 94,92% |
| Bayesian Boosting with Random Forest | 85,88% | 86,86% | 85,63% | 86,34% | 85,73% | 86,38% | 85,67% | 87,09% | 86,40% | 86,87% | 85,86% | 86,71% |
| AdaBoost with Random Forest | 83,08% | 87,68% | 83,18% | 85,78% | 83,01% | 86,01% | 83,90% | 85,72% | 82,13% | 86,23% | 83,06% | 86,28% |
| Bagging with Random Forest | 83,09% | 87,13% | 84,21% | 87,40% | 85,43% | 87,82% | 85,10% | 87,70% | 85,28% | 87,22% | 84,62% | 87,45% |
| Bayesian Boosting with Naive Bayes | 83,54% | 83,39% | 84,28% | 84,02% | 85,58% | 84,56% | 84,38% | 84,37% | 84,14% | 84,28% | 84,38% | 84,12% |
| AdaBoost with Naive Bayes | 84,86% | 84,08% | 84,76% | 83,44% | 84,80% | 84,32% | 84,40% | 84,13% | 84,86% | 84,11% | 84,74% | 84,02% |
| Bagging with Naive Bayes | 82,17% | 82,27% | 81,61% | 82,30% | 80,33% | 82,53% | 79,88% | 82,66% | 79,73% | 82,44% | 80,74% | 82,44% |
| Bayesian Boosting with K-NN | 94,69% | 94,00% | 93,65% | 93,62% | 92,81% | 93,01% | 92,45% | 92,40% | 92,33% | 91,94% | 93,19% | 92,99% |
| AdaBoost with K-NN | 94,69% | 94,00% | 93,65% | 93,62% | 92,81% | 93,01% | 92,45% | 92,40% | 92,33% | 91,94% | 93,19% | 92,99% |
| Bagging with K-NN | 94,69% | 93,94% | 93,64% | 93,59% | 92,74% | 93,12% | 92,32% | 92,36% | 92,55% | 91,85% | 93,19% | 92,97% |

Table C.1.26.: Survivability percentage AUC of unbalanced models, for rectal cancer.

| AUC - Rectal Cancer Unbalanced Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 0,958 | 0,957 | 0,967 | 0,960 | 0,971 | 0,961 | 0,971 | 0,963 | 0,972 | 0,971 | 0,968 | 0,961 |
| Voting | 0,973 | 0,953 | 0,979 | 0,964 | 0,983 | 0,968 | 0,978 | 0,961 | 0,979 | 0,962 | 0,978 | 0,962 |
| Bayesian Boosting with Decision Tree | 0,951 | 0,938 | 0,961 | 0,940 | 0,972 | 0,949 | 0,966 | 0,955 | 0,978 | 0,955 | 0,805 | 0,790 |
| AdaBoost with Decision Tree | 0,967 | 0,941 | 0,966 | 0,948 | 0,977 | 0,951 | 0,973 | 0,958 | 0,976 | 0,957 | 0,972 | 0,951 |
| Bagging with Decision Tree | 0,953 | 0,948 | 0,959 | 0,956 | 0,959 | 0,957 | 0,944 | 0,940 | 0,948 | 0,942 | 0,953 | 0,949 |
| Bayesian Boosting with Random Forest | 0,793 | 0,827 | 0,850 | 0,859 | 0,875 | 0,875 | 0,884 | 0,898 | 0,895 | 0,900 | 0,859 | 0,872 |
| AdaBoost with Random Forest | 0,744 | 0,814 | 0,816 | 0,857 | 0,832 | 0,873 | 0,869 | 0,893 | 0,858 | 0,896 | 0,824 | 0,867 |
| Bagging with Random Forest | 0,936 | 0,935 | 0,934 | 0,945 | 0,938 | 0,949 | 0,935 | 0,950 | 0,937 | 0,951 | 0,936 | 0,946 |
| Bayesian Boosting with Naive Bayes | 0,888 | 0,875 | 0,904 | 0,889 | 0,910 | 0,900 | 0,910 | 0,903 | 0,907 | 0,900 | 0,904 | 0,893 |
| AdaBoost with Naive Bayes | 0,883 | 0,870 | 0,901 | 0,891 | 0,908 | 0,898 | 0,914 | 0,901 | 0,914 | 0,901 | 0,904 | 0,892 |
| Bagging with Naive Bayes | 0,872 | 0,875 | 0,884 | 0,889 | 0,896 | 0,902 | 0,898 | 0,905 | 0,899 | 0,905 | 0,890 | 0,895 |
| Bayesian Boosting with K-NN | 0,830 | 0,859 | 0,805 | 0,886 | 0,806 | 0,888 | 0,889 | 0,890 | 0,855 | 0,891 | 0,837 | 0,883 |
| AdaBoost with K-NN | 0,867 | 0,859 | 0,882 | 0,886 | 0,883 | 0,888 | 0,889 | 0,890 | 0,893 | 0,891 | 0,883 | 0,883 |
| Bagging with K-NN | 0,892 | 0,887 | 0,900 | 0,906 | 0,905 | 0,913 | 0,908 | 0,915 | 0,916 | 0,913 | 0,904 | 0,907 |

Table C.1.27.: F-measure performance of unbalanced models, for rectal cancer.

| F-Measure - Rectal Cancer Unbalanced Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 78,16% | 77,63% | 83,92% | 84,16% | 85,90% | 85,96% | 96,40% | 95,97% | 96,42% | 96,40% | 88,16% | 88,03% |
| Voting | 93,70% | 90,01% | 95,28% | 92,58% | 95,97% | 93,67% | 97,67% | 96,47% | 97,64% | 96,12% | 96,05% | 93,77% |
| Bayesian Boosting with Decision Tree | 92,93% | 90,22% | 94,57% | 92,61% | 95,57% | 93,52% | 97,41% | 96,29% | 97,18% | 96,24% | 79,61% | 78,15% |
| AdaBoost with Decision Tree | 92,93% | 90,22% | 94,57% | 92,61% | 95,57% | 93,52% | 97,41% | 96,29% | 97,18% | 96,24% | 95,54% | 93,78% |
| Bagging with Decision Tree | 92,40% | 89,78% | 93,88% | 91,63% | 94,71% | 92,73% | 96,72% | 95,76% | 96,74% | 95,84% | 94,89% | 93,15% |
| Bayesian Boosting with Random Forest | 63,93% | 68,40% | 74,24% | 76,26% | 77,95% | 79,48% | 88,81% | 89,91% | 89,13% | 89,48% | 78,81% | 80,71% |
| AdaBoost with Random Forest | 51,36% | 70,80% | 67,61% | 75,30% | 71,91% | 78,71% | 87,77% | 89,06% | 86,26% | 89,06% | 72,98% | 80,59% |
| Bagging with Random Forest | 49,45% | 68,34% | 69,38% | 77,93% | 76,51% | 81,66% | 88,85% | 90,46% | 88,67% | 89,89% | 74,57% | 81,65% |
| Bayesian Boosting with Naive Bayes | 64,93% | 61,60% | 75,73% | 74,39% | 79,69% | 78,29% | 87,61% | 87,49% | 87,19% | 87,25% | 79,03% | 77,80% |
| AdaBoost with Naive Bayes | 67,80% | 63,52% | 75,18% | 73,69% | 78,39% | 77,61% | 87,29% | 87,32% | 87,40% | 87,07% | 79,21% | 77,84% |
| Bagging with Naive Bayes | 65,61% | 62,18% | 73,09% | 73,07% | 74,93% | 76,26% | 82,47% | 85,71% | 81,82% | 85,16% | 75,58% | 76,48% |
| Bayesian Boosting with K-NN | 77,51% | 75,14% | 81,60% | 81,75% | 82,77% | 83,38% | 95,06% | 95,02% | 94,86% | 94,57% | 86,36% | 85,97% |
| AdaBoost with K-NN | 77,51% | 75,14% | 81,60% | 81,75% | 82,77% | 83,38% | 95,06% | 95,02% | 94,86% | 94,57% | 86,36% | 85,97% |
| Bagging with K-NN | 77,51% | 74,79% | 83,87% | 81,59% | 82,64% | 85,57% | 94,97% | 95,00% | 95,00% | 94,53% | 86,80% | 86,30% |

Table C.1.28.: Percentage of wrongly classified cases of unbalanced models, for rectal cancer.

| Wrongly Classified Cases - Rectal Cancer Unbalanced Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 5,46% | 5,23% | 5,62% | 4,68% | 5,46% | 5,07% | 4,37% | 6,32% | 4,99% | 6,01% | 5,18% | 5,46% |
| Voting | 12,64% | 12,40% | 14,27% | 16,54% | 15,60% | 17,39% | 20,20% | 20,12% | 22,07% | 20,28% | 16,96% | 17,35% |
| Bayesian Boosting with Decision Tree | 16,30% | 16,07% | 19,34% | 17,78% | 17,24% | 19,03% | 35,96% | 27,54% | 29,41% | 28,78% | 23,65% | 21,84% |
| AdaBoost with Decision Tree | 16,30% | 16,07% | 19,34% | 17,78% | 16,46% | 19,03% | 35,96% | 27,54% | 29,41% | 28,71% | 23,49% | 21,83% |
| Bagging with Decision Tree | 13,10% | 15,52% | 16,61% | 17,63% | 16,61% | 20,12% | 46,88% | 31,51% | 58,19% | 33,07% | 30,28% | 23,57% |
| Bayesian Boosting with Random Forest | 11,78% | 11,62% | 15,44% | 13,96% | 14,98% | 15,05% | 18,41% | 18,02% | 21,92% | 17,55% | 16,51% | 15,24% |
| AdaBoost with Random Forest | 11,86% | 11,86% | 16,69% | 13,10% | 17,94% | 17,94% | 16,77% | 17,86% | 17,16% | 18,41% | 16,08% | 15,83% |
| Bagging with Random Forest | 11,39% | 11,00% | 13,49% | 13,73% | 16,69% | 15,05% | 15,76% | 16,46% | 17,08% | 16,93% | 14,88% | 14,63% |
| Bayesian Boosting with Naive Bayes | 13,42% | 13,81% | 15,76% | 15,37% | 15,05% | 15,99% | 17,55% | 17,63% | 16,93% | 17,55% | 13,12% | 13,39% |
| AdaBoost with Naive Bayes | 20,90% | 13,42% | 19,34% | 15,99% | 18,17% | 15,99% | 19,42% | 17,08% | 20,20% | 17,55% | 19,61% | 16,01% |
| Bagging with Naive Bayes | 15,37% | 14,43% | 16,15% | 16,07% | 18,72% | 15,99% | 20,28% | 17,71% | 19,73% | 18,41% | 18,05% | 16,52% |
| Bayesian Boosting with K-NN | 5,30% | 5,54% | 5,62% | 5,46% | 6,08% | 5,69% | 6,16% | 6,94% | 6,40% | 6,94% | 5,91% | 6,12% |
| AdaBoost with K-NN | 5,30% | 5,54% | 5,62% | 5,46% | 6,08% | 5,69% | 6,16% | 6,94% | 6,40% | 6,94% | 5,91% | 6,12% |
| Bagging with K-NN | 5,30% | 5,54% | 5,62% | 5,46% | 6,08% | 5,69% | 6,16% | 6,94% | 6,40% | 6,94% | 5,91% | 6,12% |

*Hybrid Models*

Table C.1.29.: Survivability percentage accuracy of hybrid models, for rectal cancer.

| Accuracy - Rectal Cancer Hybrid Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 97,21% | 96,37% | 97,61% | 96,12% | 97,16% | 95,66% | 97,18% | 95,17% | 96,92% | 95,00% | 97,22% | 95,66% |
| Voting | 97,16% | 95,49% | 97,14% | 95,84% | 96,54% | 95,42% | 96,49% | 94,92% | 96,32% | 94,38% | 96,73% | 95,21% |
| Bayesian Boosting with Decision Tree | 97,36% | 95,74% | 96,67% | 95,73% | 96,13% | 95,06% | 95,74% | 95,11% | 96,06% | 94,43% | 80,33% | 79,35% |
| AdaBoost with Decision Tree | 97,36% | 95,74% | 96,67% | 95,73% | 96,13% | 95,06% | 95,74% | 95,11% | 95,93% | 94,43% | 96,37% | 95,21% |
| Bagging with Decision Tree | 95,34% | 93,79% | 94,76% | 93,60% | 94,44% | 93,20% | 94,67% | 93,02% | 94,62% | 91,88% | 94,77% | 93,10% |
| Bayesian Boosting with Random Forest | 84,83% | 87,54% | 83,55% | 86,36% | 84,44% | 86,16% | 83,20% | 85,14% | 83,29% | 84,78% | 83,86% | 86,00% |
| AdaBoost with Random Forest | 83,71% | 88,95% | 83,04% | 85,93% | 83,67% | 85,52% | 81,32% | 84,58% | 81,80% | 83,76% | 82,71% | 85,75% |
| Bagging with Random Forest | 86,29% | 89,88% | 85,61% | 87,53% | 85,97% | 87,59% | 84,74% | 86,37% | 84,14% | 86,81% | 85,35% | 87,64% |
| Bayesian Boosting with Naive Bayes | 83,64% | 82,66% | 82,31% | 82,71% | 82,72% | 82,04% | 81,15% | 81,32% | 81,12% | 80,55% | 82,19% | 81,86% |
| AdaBoost with Naive Bayes | 83,69% | 82,99% | 82,42% | 82,58% | 83,03% | 82,14% | 81,06% | 81,32% | 81,20% | 80,56% | 82,28% | 81,92% |
| Bagging with Naive Bayes | 81,72% | 82,48% | 81,89% | 82,79% | 82,71% | 81,97% | 80,84% | 81,31% | 80,08% | 80,42% | 81,45% | 81,79% |
| Bayesian Boosting with K-NN | 97,03% | 93,98% | 96,19% | 93,44% | 95,48% | 93,10% | 95,41% | 92,27% | 95,09% | 91,86% | 95,84% | 92,93% |
| AdaBoost with K-NN | 97,03% | 93,98% | 96,19% | 93,44% | 95,48% | 93,10% | 95,41% | 92,27% | 95,09% | 91,86% | 95,84% | 92,93% |
| Bagging with K-NN | 97,03% | 93,98% | 96,20% | 93,45% | 95,47% | 92,95% | 95,73% | 91,96% | 95,08% | 91,95% | 95,90% | 92,86% |

Table C.1.30.: Survivability percentage AUC of hybrid models, for rectal cancer.

| AUC - Rectal Cancer Hybrid Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 0,987 | 0,974 | 0,985 | 0,979 | 0,985 | 0,978 | 0,984 | 0,977 | 0,984 | 0,974 | 0,985 | 0,976 |
| Voting | 0,985 | 0,971 | 0,982 | 0,975 | 0,980 | 0,971 | 0,978 | 0,968 | 0,977 | 0,962 | 0,980 | 0,969 |
| Bayesian Boosting with Decision Tree | 0,500 | 0,958 | 0,958 | 0,953 | 0,950 | 0,953 | 0,813 | 0,957 | 0,953 | 0,947 | 0,696 | 0,795 |
| AdaBoost with Decision Tree | 0,933 | 0,964 | 0,969 | 0,967 | 0,968 | 0,959 | 0,964 | 0,958 | 0,968 | 0,951 | 0,960 | 0,960 |
| Bagging with Decision Tree | 0,984 | 0,978 | 0,973 | 0,961 | 0,966 | 0,949 | 0,964 | 0,945 | 0,961 | 0,939 | 0,970 | 0,954 |
| Bayesian Boosting with Random Forest | 0,909 | 0,920 | 0,908 | 0,915 | 0,910 | 0,919 | 0,901 | 0,911 | 0,900 | 0,903 | 0,906 | 0,914 |
| AdaBoost with Random Forest | 0,902 | 0,919 | 0,902 | 0,913 | 0,903 | 0,912 | 0,887 | 0,900 | 0,890 | 0,895 | 0,897 | 0,908 |
| Bagging with Random Forest | 0,921 | 0,953 | 0,928 | 0,948 | 0,934 | 0,948 | 0,927 | 0,940 | 0,921 | 0,938 | 0,926 | 0,945 |
| Bayesian Boosting with Naive Bayes | 0,892 | 0,899 | 0,900 | 0,897 | 0,907 | 0,896 | 0,893 | 0,880 | 0,892 | 0,877 | 0,897 | 0,890 |
| AdaBoost with Naive Bayes | 0,893 | 0,898 | 0,905 | 0,903 | 0,908 | 0,897 | 0,893 | 0,882 | 0,895 | 0,882 | 0,899 | 0,892 |
| Bagging with Naive Bayes | 0,873 | 0,901 | 0,896 | 0,901 | 0,898 | 0,900 | 0,884 | 0,888 | 0,876 | 0,886 | 0,885 | 0,895 |
| Bayesian Boosting with K-NN | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 |
| AdaBoost with K-NN | 0,500 | 0,940 | 0,962 | 0,934 | 0,955 | 0,931 | 0,954 | 0,923 | 0,951 | 0,919 | 0,864 | 0,929 |
| Bagging with K-NN | 0,971 | 0,942 | 0,964 | 0,939 | 0,959 | 0,937 | 0,962 | 0,930 | 0,958 | 0,929 | 0,963 | 0,935 |

Table C.1.31.: F-measure performance of hybrid models, for rectal cancer.

| F-Measure - Rectal Cancer Hybrid Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 97,14% | 96,26% | 97,59% | 96,01% | 97,15% | 95,54% | 97,17% | 95,04% | 96,93% | 94,88% | 97,20% | 95,55% |
| Voting | 97,09% | 95,32% | 97,09% | 95,71% | 96,48% | 95,27% | 96,42% | 94,72% | 96,27% | 94,17% | 96,67% | 95,04% |
| Bayesian Boosting with Decision Tree | 97,30% | 95,58% | 96,59% | 95,60% | 96,04% | 94,90% | 95,65% | 94,98% | 96,02% | 94,26% | 80,27% | 79,22% |
| AdaBoost with Decision Tree | 97,30% | 95,58% | 96,59% | 95,60% | 96,04% | 94,90% | 95,65% | 94,98% | 95,86% | 94,26% | 96,29% | 95,06% |
| Bagging with Decision Tree | 95,12% | 93,44% | 94,52% | 93,27% | 94,19% | 92,82% | 94,44% | 92,66% | 94,40% | 91,39% | 94,54% | 92,72% |
| Bayesian Boosting with Random Forest | 84,78% | 87,45% | 83,87% | 86,50% | 84,72% | 86,36% | 83,75% | 85,12% | 83,81% | 84,80% | 84,19% | 86,05% |
| AdaBoost with Random Forest | 83,46% | 88,69% | 83,43% | 86,09% | 84,57% | 85,76% | 82,35% | 84,58% | 82,83% | 83,86% | 83,33% | 85,79% |
| Bagging with Random Forest | 86,26% | 89,78% | 86,10% | 87,78% | 86,78% | 87,94% | 85,61% | 86,65% | 85,22% | 87,02% | 85,99% | 87,83% |
| Bayesian Boosting with Naive Bayes | 83,91% | 83,08% | 83,15% | 83,26% | 83,20% | 82,64% | 81,32% | 82,33% | 80,31% | 81,67% | 82,38% | 82,60% |
| AdaBoost with Naive Bayes | 83,91% | 83,43% | 83,31% | 83,20% | 83,46% | 82,66% | 81,16% | 82,33% | 80,48% | 81,61% | 82,46% | 82,65% |
| Bagging with Naive Bayes | 82,55% | 83,04% | 83,02% | 83,37% | 83,43% | 82,87% | 81,37% | 82,33% | 80,14% | 81,54% | 82,10% | 82,63% |
| Bayesian Boosting with K-NN | 96,95% | 93,61% | 96,09% | 93,02% | 95,37% | 92,67% | 95,31% | 91,73% | 95,00% | 91,26% | 95,75% | 92,46% |
| AdaBoost with K-NN | 96,95% | 93,61% | 96,09% | 93,02% | 95,37% | 92,67% | 95,31% | 91,73% | 95,00% | 91,26% | 95,75% | 92,46% |
| Bagging with K-NN | 96,95% | 93,62% | 96,10% | 93,04% | 95,36% | 92,50% | 95,64% | 91,37% | 94,98% | 91,36% | 95,81% | 92,38% |

Table C.1.32.: Percentage of wrongly classified cases of hybrid models, for rectal cancer.

| Wrongly Classified Cases - Rectal Cancer Hybrid Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 39,63% | 35,96% | 35,65% | 29,25% | 57,33% | 41,58% | 53,12% | 32,45% | 55,93% | 39,24% | 48,33% | 35,69% |
| Voting | 36,04% | 56,47% | 42,98% | 45,71% | 37,52% | 63,42% | 32,68% | 59,83% | 38,38% | 61,86% | 37,52% | 57,46% |
| Bayesian Boosting with Decision Tree | 63,96% | 51,95% | 79,64% | 45,87% | 44,77% | 63,57% | 37,36% | 58,50% | 44,46% | 63,18% | 54,04% | 56,61% |
| AdaBoost with Decision Tree | 63,96% | 51,95% | 79,64% | 45,87% | 44,77% | 63,57% | 37,36% | 58,50% | 44,46% | 63,18% | 54,04% | 56,61% |
| Bagging with Decision Tree | 79,10% | 65,37% | 83,00% | 61,86% | 47,04% | 69,34% | 74,34% | 67,32% | 45,09% | 64,04% | 65,71% | 65,59% |
| Bayesian Boosting with Random Forest | 40,41% | 55,38% | 72,31% | 41,81% | 75,12% | 63,03% | 58,97% | 59,05% | 67,86% | 63,96% | 62,93% | 56,65% |
| AdaBoost with Random Forest | 56,79% | 69,73% | 76,99% | 47,82% | 68,10% | 66,85% | 62,01% | 55,38% | 48,44% | 62,87% | 62,46% | 60,53% |
| Bagging with Random Forest | 39,94% | 57,49% | 41,26% | 40,80% | 73,95% | 62,09% | 70,51% | 58,58% | 70,51% | 62,56% | 59,24% | 56,30% |
| Bayesian Boosting with Naive Bayes | 55,46% | 39,16% | 60,45% | 39,16% | 76,76% | 60,30% | 74,34% | 54,84% | 72,78% | 58,74% | 56,63% | 42,03% |
| AdaBoost with Naive Bayes | 53,90% | 39,70% | 60,84% | 37,75% | 76,68% | 56,32% | 74,34% | 54,84% | 72,78% | 58,74% | 67,71% | 49,47% |
| Bagging with Naive Bayes | 47,50% | 43,60% | 50,08% | 39,31% | 76,68% | 58,35% | 74,41% | 54,91% | 72,78% | 58,74% | 64,29% | 50,98% |
| Bayesian Boosting with K-NN | 16,85% | 45,94% | 19,03% | 45,01% | 17,47% | 60,45% | 19,27% | 57,72% | 20,98% | 56,86% | 18,72% | 53,20% |
| AdaBoost with K-NN | 16,85% | 45,94% | 19,03% | 45,01% | 17,47% | 60,45% | 19,27% | 57,72% | 20,98% | 56,86% | 18,72% | 53,20% |
| Bagging with K-NN | 16,85% | 45,94% | 19,03% | 45,01% | 17,47% | 60,45% | 19,27% | 57,72% | 20,98% | 56,86% | 18,72% | 53,20% |

*Oversampled Models*

Table C.1.33.: Survivability percentage accuracy of oversampled models, for rectal cancer.

| Accuracy - Rectal Cancer Oversampled Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 99,00% | 97,18% | 98,73% | 96,87% | 98,55% | 96,69% | 98,34% | 96,24% | 98,00% | 96,06% | 98,52% | 96,61% |
| Voting | 98,30% | 96,79% | 98,24% | 96,66% | 98,16% | 96,62% | 97,94% | 96,07% | 97,75% | 95,68% | 98,08% | 96,36% |
| Bayesian Boosting with Decision Tree | 98,34% | 97,04% | 98,01% | 96,71% | 97,80% | 96,64% | 97,29% | 96,08% | 97,30% | 95,81% | 81,46% | 80,38% |
| AdaBoost with Decision Tree | 98,34% | 97,04% | 98,01% | 96,71% | 97,80% | 96,64% | 97,23% | 96,08% | 97,27% | 95,81% | 97,73% | 96,46% |
| Bagging with Decision Tree | 97,20% | 95,83% | 96,75% | 95,36% | 96,49% | 95,29% | 96,59% | 94,24% | 96,09% | 94,04% | 96,62% | 94,95% |
| Bayesian Boosting with Random Forest | 84,72% | 88,22% | 84,97% | 86,32% | 84,76% | 85,32% | 83,67% | 85,00% | 83,61% | 84,68% | 84,35% | 85,91% |
| AdaBoost with Random Forest | 83,11% | 87,81% | 84,29% | 86,50% | 83,53% | 85,73% | 82,14% | 84,28% | 83,73% | 84,28% | 83,36% | 85,72% |
| Bagging with Random Forest | 86,58% | 89,98% | 85,68% | 88,01% | 86,21% | 87,79% | 84,76% | 86,88% | 84,50% | 86,54% | 85,55% | 87,84% |
| Bayesian Boosting with Naive Bayes | 83,70% | 83,11% | 82,18% | 82,58% | 82,87% | 82,82% | 82,18% | 80,70% | 81,09% | 80,77% | 82,40% | 82,00% |
| AdaBoost with Naive Bayes | 83,18% | 82,60% | 82,43% | 82,59% | 82,95% | 82,87% | 82,40% | 80,89% | 81,18% | 80,60% | 82,43% | 81,91% |
| Bagging with Naive Bayes | 81,73% | 82,65% | 81,49% | 82,82% | 82,69% | 82,72% | 80,82% | 80,78% | 80,15% | 80,76% | 81,38% | 81,95% |
| Bayesian Boosting with K-NN | 98,27% | 95,63% | 97,82% | 95,01% | 97,26% | 94,63% | 97,09% | 93,86% | 96,98% | 93,13% | 97,48% | 94,45% |
| AdaBoost with K-NN | 98,20% | 95,65% | 97,77% | 94,96% | 97,18% | 94,66% | 97,14% | 93,68% | 96,91% | 93,32% | 97,44% | 94,45% |
| Bagging with K-NN | 98,27% | 95,59% | 97,86% | 95,00% | 97,31% | 94,55% | 97,03% | 93,80% | 96,96% | 93,21% | 97,49% | 94,43% |

Table C.1.34.: Survivability percentage AUC of oversampled models, for rectal cancer.

| AUC - Rectal Cancer Oversampled Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 0,993 | 0,980 | 0,993 | 0,982 | 0,992 | 0,982 | 0,990 | 0,982 | 0,990 | 0,982 | 0,992 | 0,982 |
| Voting | 0,991 | 0,980 | 0,989 | 0,980 | 0,989 | 0,978 | 0,988 | 0,978 | 0,986 | 0,975 | 0,989 | 0,978 |
| Bayesian Boosting with Decision Tree | 0,988 | 0,974 | 0,926 | 0,963 | 0,969 | 0,969 | 0,961 | 0,958 | 0,967 | 0,955 | 0,802 | 0,803 |
| AdaBoost with Decision Tree | 0,984 | 0,977 | 0,883 | 0,971 | 0,984 | 0,968 | 0,980 | 0,965 | 0,981 | 0,964 | 0,962 | 0,969 |
| Bagging with Decision Tree | 0,995 | 0,990 | 0,989 | 0,982 | 0,984 | 0,974 | 0,982 | 0,967 | 0,979 | 0,970 | 0,986 | 0,977 |
| Bayesian Boosting with Random Forest | 0,901 | 0,926 | 0,914 | 0,917 | 0,917 | 0,913 | 0,907 | 0,912 | 0,903 | 0,908 | 0,908 | 0,915 |
| AdaBoost with Random Forest | 0,904 | 0,913 | 0,905 | 0,911 | 0,901 | 0,917 | 0,885 | 0,908 | 0,898 | 0,905 | 0,899 | 0,911 |
| Bagging with Random Forest | 0,920 | 0,957 | 0,932 | 0,948 | 0,938 | 0,949 | 0,928 | 0,940 | 0,924 | 0,940 | 0,928 | 0,947 |
| Bayesian Boosting with Naive Bayes | 0,894 | 0,900 | 0,900 | 0,904 | 0,912 | 0,900 | 0,897 | 0,882 | 0,892 | 0,878 | 0,899 | 0,893 |
| AdaBoost with Naive Bayes | 0,894 | 0,898 | 0,902 | 0,904 | 0,911 | 0,900 | 0,895 | 0,885 | 0,894 | 0,883 | 0,899 | 0,894 |
| Bagging with Naive Bayes | 0,873 | 0,904 | 0,893 | 0,901 | 0,901 | 0,901 | 0,881 | 0,887 | 0,877 | 0,886 | 0,885 | 0,896 |
| Bayesian Boosting with K-NN | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 |
| AdaBoost with K-NN | 0,982 | 0,957 | 0,978 | 0,950 | 0,972 | 0,947 | 0,971 | 0,937 | 0,969 | 0,933 | 0,974 | 0,945 |
| Bagging with K-NN | 0,983 | 0,956 | 0,979 | 0,951 | 0,975 | 0,949 | 0,972 | 0,940 | 0,972 | 0,937 | 0,976 | 0,947 |

Table C.1.35.: F-measure performance of oversampled models, for rectal cancer.

| F-Measure - Rectal Cancer Oversampled Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 99,00% | 97,11% | 98,72% | 96,79% | 98,54% | 96,61% | 98,33% | 96,14% | 97,99% | 95,97% | 98,52% | 96,53% |
| Voting | 98,27% | 96,70% | 98,21% | 96,56% | 98,13% | 96,52% | 97,91% | 95,95% | 97,71% | 95,53% | 98,05% | 96,25% |
| Bayesian Boosting with Decision Tree | 98,31% | 96,97% | 97,97% | 96,62% | 97,76% | 96,56% | 97,24% | 95,98% | 97,25% | 95,70% | 81,42% | 80,30% |
| AdaBoost with Decision Tree | 98,31% | 96,97% | 97,97% | 96,62% | 97,76% | 96,56% | 97,17% | 95,98% | 97,22% | 95,70% | 97,69% | 96,37% |
| Bagging with Decision Tree | 97,12% | 95,68% | 96,65% | 95,18% | 96,38% | 95,11% | 96,49% | 93,98% | 95,96% | 93,77% | 96,52% | 94,74% |
| Bayesian Boosting with Random Forest | 84,82% | 88,09% | 85,23% | 86,62% | 85,16% | 85,51% | 84,25% | 85,13% | 83,74% | 84,82% | 84,64% | 86,03% |
| AdaBoost with Random Forest | 82,36% | 87,73% | 84,58% | 86,71% | 84,15% | 85,91% | 83,25% | 84,26% | 84,33% | 84,38% | 83,73% | 85,80% |
| Bagging with Random Forest | 86,63% | 89,91% | 86,37% | 88,28% | 87,04% | 88,14% | 85,74% | 87,04% | 85,39% | 86,76% | 86,23% | 88,03% |
| Bayesian Boosting with Naive Bayes | 84,03% | 83,34% | 83,02% | 83,09% | 83,53% | 83,30% | 81,67% | 81,73% | 80,27% | 81,77% | 82,50% | 82,65% |
| AdaBoost with Naive Bayes | 83,72% | 83,10% | 83,29% | 83,17% | 83,54% | 83,51% | 82,00% | 81,84% | 80,49% | 81,55% | 82,61% | 82,63% |
| Bagging with Naive Bayes | 82,61% | 83,24% | 82,72% | 83,42% | 83,49% | 83,50% | 80,89% | 81,81% | 80,14% | 81,77% | 81,97% | 82,75% |
| Bayesian Boosting with K-NN | 98,16% | 95,46% | 97,73% | 94,70% | 97,12% | 94,38% | 97,08% | 93,29% | 96,85% | 92,88% | 97,39% | 94,14% |
| AdaBoost with K-NN | 98,16% | 95,46% | 97,73% | 94,70% | 97,12% | 94,38% | 97,08% | 93,29% | 96,85% | 92,88% | 97,39% | 94,14% |
| Bagging with K-NN | 98,24% | 95,39% | 97,82% | 94,75% | 97,26% | 94,25% | 96,97% | 93,42% | 96,90% | 92,76% | 97,44% | 94,11% |

Table C.1.36.: Percentage of wrongly classified cases of oversampled models, for rectal cancer.

| Wrongly Classified Cases - Rectal Cancer Oversampled Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 18,95% | 32,76% | 36,27% | 24,49% | 45,09% | 32,92% | 57,18% | 30,73% | 53,74% | 40,17% | 42,25% | 32,22% |
| Voting | 53,82% | 51,56% | 40,48% | 40,41% | 33,78% | 54,52% | 30,34% | 71,29% | 40,41% | 59,98% | 39,77% | 55,55% |
| Bayesian Boosting with Decision Tree | 61,93% | 50,94% | 79,56% | 42,75% | 39,16% | 53,98% | 38,46% | 64,12% | 44,54% | 61,78% | 52,73% | 54,71% |
| AdaBoost with Decision Tree | 61,93% | 50,94% | 79,56% | 42,75% | 39,16% | 53,98% | 38,46% | 64,12% | 44,54% | 61,78% | 52,73% | 54,71% |
| Bagging with Decision Tree | 74,65% | 58,42% | 81,51% | 63,81% | 48,60% | 58,27% | 50,47% | 67,55% | 65,83% | 69,73% | 64,21% | 63,56% |
| Bayesian Boosting with Random Forest | 46,65% | 57,18% | 63,34% | 44,54% | 75,66% | 51,40% | 66,77% | 66,30% | 67,78% | 65,60% | 64,04% | 57,00% |
| AdaBoost with Random Forest | 58,58% | 45,40% | 26,91% | 37,52% | 52,57% | 49,92% | 44,70% | 51,40% | 66,54% | 56,63% | 49,86% | 48,17% |
| Bagging with Random Forest | 37,75% | 61,39% | 42,36% | 44,23% | 67,71% | 55,38% | 63,73% | 68,02% | 70,98% | 65,21% | 56,51% | 58,85% |
| Bayesian Boosting with Naive Bayes | 57,02% | 41,81% | 56,08% | 41,58% | 76,68% | 50,70% | 74,34% | 61,23% | 72,78% | 60,06% | 56,15% | 42,56% |
| AdaBoost with Naive Bayes | 52,73% | 39,94% | 56,63% | 39,55% | 76,68% | 43,37% | 74,34% | 55,07% | 72,78% | 60,06% | 66,63% | 47,60% |
| Bagging with Naive Bayes | 48,75% | 44,54% | 43,21% | 39,86% | 76,68% | 50,39% | 74,10% | 61,00% | 72,78% | 60,06% | 63,10% | 51,17% |
| Bayesian Boosting with K-NN | 15,13% | 43,76% | 14,82% | 44,07% | 17,47% | 51,79% | 17,32% | 59,59% | 20,12% | 55,85% | 16,97% | 51,01% |
| AdaBoost with K-NN | 15,13% | 43,76% | 14,82% | 44,07% | 17,47% | 51,79% | 17,32% | 59,59% | 20,12% | 55,85% | 16,97% | 51,01% |
| Bagging with K-NN | 15,13% | 43,76% | 14,82% | 44,07% | 17,47% | 51,79% | 17,32% | 59,59% | 20,12% | 55,85% | 16,97% | 51,01% |

*Undersampled Models*

Table C.1.37.: Survivability percentage accuracy of undersampled models, for rectal cancer.

| Accuracy - Rectal Cancer Undersampled Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 90,56% | 88,90% | 88,70% | 89,42% | 90,24% | 89,61% | 89,96% | 90,17% | 89,98% | 90,47% | 89,89% | 89,71% |
| Voting | 90,05% | 89,37% | 90,15% | 89,42% | 90,62% | 89,33% | 90,58% | 90,12% | 90,78% | 90,13% | 90,44% | 89,67% |
| Bayesian Boosting with Decision Tree | 88,58% | 87,35% | 87,48% | 88,06% | 88,94% | 88,52% | 89,15% | 88,95% | 89,05% | 89,25% | 73,82% | 73,69% |
| AdaBoost with Decision Tree | 88,58% | 87,35% | 87,48% | 88,06% | 88,86% | 88,52% | 89,15% | 88,95% | 88,84% | 89,25% | 88,58% | 88,43% |
| Bagging with Decision Tree | 87,20% | 85,44% | 86,17% | 85,83% | 87,46% | 85,76% | 87,37% | 86,62% | 88,07% | 87,31% | 87,25% | 86,19% |
| Bayesian Boosting with Random Forest | 82,73% | 84,28% | 83,03% | 83,79% | 82,67% | 84,12% | 81,26% | 83,50% | 81,84% | 83,01% | 82,31% | 83,74% |
| AdaBoost with Random Forest | 80,89% | 84,90% | 80,72% | 83,87% | 81,15% | 82,65% | 81,13% | 83,09% | 81,00% | 82,48% | 80,98% | 83,40% |
| Bagging with Random Forest | 84,10% | 87,10% | 84,28% | 85,58% | 83,50% | 85,48% | 83,61% | 84,83% | 82,39% | 85,08% | 83,58% | 85,61% |
| Bayesian Boosting with Naive Bayes | 82,05% | 81,54% | 82,42% | 82,96% | 82,62% | 82,28% | 81,80% | 81,10% | 80,91% | 80,30% | 81,96% | 81,64% |
| AdaBoost with Naive Bayes | 82,66% | 81,08% | 82,15% | 82,76% | 82,95% | 82,26% | 81,73% | 80,99% | 81,06% | 80,71% | 82,11% | 81,56% |
| Bagging with Naive Bayes | 81,22% | 81,98% | 81,51% | 82,84% | 82,44% | 81,92% | 80,81% | 81,21% | 80,17% | 80,37% | 81,23% | 81,66% |
| Bayesian Boosting with K-NN | 85,80% | 84,32% | 86,67% | 85,46% | 86,74% | 84,12% | 87,39% | 85,98% | 86,98% | 85,73% | 86,72% | 85,12% |
| AdaBoost with K-NN | 85,80% | 84,32% | 86,67% | 85,46% | 86,74% | 84,12% | 87,39% | 85,98% | 86,98% | 85,73% | 86,72% | 85,12% |
| Bagging with K-NN | 85,80% | 84,54% | 86,74% | 84,82% | 86,92% | 83,84% | 87,12% | 86,28% | 87,48% | 86,01% | 86,81% | 85,10% |

Table C.1.38.: Survivability percentage AUC of undersampled models, for rectal cancer.

| AUC - Rectal Cancer Undersampled Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 0,945 | 0,950 | 0,953 | 0,953 | 0,956 | 0,951 | 0,959 | 0,954 | 0,958 | 0,954 | 0,954 | 0,952 |
| Voting | 0,925 | 0,920 | 0,932 | 0,927 | 0,938 | 0,928 | 0,941 | 0,932 | 0,939 | 0,936 | 0,935 | 0,929 |
| Bayesian Boosting with Decision Tree | 0,500 | 0,863 | 0,781 | 0,903 | 0,611 | 0,896 | 0,897 | 0,898 | 0,764 | 0,889 | 0,439 | 0,748 |
| AdaBoost with Decision Tree | 0,125 | 0,888 | 0,736 | 0,891 | 0,120 | 0,894 | 0,911 | 0,915 | 0,742 | 0,898 | 0,527 | 0,897 |
| Bagging with Decision Tree | 0,789 | 0,723 | 0,842 | 0,780 | 0,860 | 0,821 | 0,875 | 0,812 | 0,880 | 0,830 | 0,849 | 0,793 |
| Bayesian Boosting with Random Forest | 0,894 | 0,895 | 0,888 | 0,897 | 0,900 | 0,897 | 0,883 | 0,892 | 0,882 | 0,891 | 0,889 | 0,894 |
| AdaBoost with Random Forest | 0,878 | 0,894 | 0,884 | 0,892 | 0,874 | 0,891 | 0,877 | 0,888 | 0,880 | 0,883 | 0,879 | 0,890 |
| Bagging with Random Forest | 0,904 | 0,938 | 0,921 | 0,936 | 0,918 | 0,934 | 0,919 | 0,931 | 0,915 | 0,928 | 0,915 | 0,933 |
| Bayesian Boosting with Naive Bayes | 0,885 | 0,885 | 0,895 | 0,892 | 0,899 | 0,890 | 0,893 | 0,876 | 0,891 | 0,880 | 0,893 | 0,885 |
| AdaBoost with Naive Bayes | 0,885 | 0,883 | 0,897 | 0,895 | 0,900 | 0,891 | 0,893 | 0,880 | 0,890 | 0,883 | 0,893 | 0,886 |
| Bagging with Naive Bayes | 0,869 | 0,895 | 0,894 | 0,898 | 0,892 | 0,894 | 0,881 | 0,889 | 0,875 | 0,883 | 0,882 | 0,892 |
| Bayesian Boosting with K-NN | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 |
| AdaBoost with K-NN | 0,500 | 0,843 | 0,794 | 0,855 | 0,500 | 0,841 | 0,874 | 0,860 | 0,794 | 0,857 | 0,692 | 0,851 |
| Bagging with K-NN | 0,891 | 0,882 | 0,893 | 0,886 | 0,899 | 0,886 | 0,901 | 0,895 | 0,903 | 0,888 | 0,897 | 0,887 |

Table C.1.39.: F-measure performance of undersampled models, for rectal cancer.

| F-Measure - Rectal Cancer Undersampled Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 90,53% | 88,92% | 89,25% | 89,30% | 90,15% | 89,55% | 90,31% | 90,02% | 90,44% | 90,28% | 90,14% | 89,61% |
| Voting | 90,00% | 88,95% | 89,91% | 89,01% | 90,43% | 88,85% | 90,40% | 89,71% | 90,65% | 89,74% | 90,28% | 89,25% |
| Bayesian Boosting with Decision Tree | 88,56% | 87,09% | 87,37% | 87,78% | 88,79% | 88,24% | 89,06% | 88,77% | 88,72% | 89,03% | 73,75% | 73,48% |
| AdaBoost with Decision Tree | 88,56% | 87,09% | 87,37% | 87,78% | 88,79% | 88,24% | 89,06% | 88,77% | 88,72% | 89,03% | 88,50% | 88,18% |
| Bagging with Decision Tree | 86,14% | 84,02% | 85,04% | 84,80% | 86,64% | 84,67% | 86,51% | 85,74% | 87,34% | 86,54% | 86,33% | 85,15% |
| Bayesian Boosting with Random Forest | 82,03% | 83,92% | 82,49% | 83,07% | 82,04% | 83,43% | 80,71% | 82,71% | 80,94% | 82,64% | 81,64% | 83,15% |
| AdaBoost with Random Forest | 80,67% | 84,44% | 79,09% | 83,01% | 79,62% | 81,57% | 79,96% | 82,47% | 79,61% | 81,84% | 79,79% | 82,66% |
| Bagging with Random Forest | 83,73% | 86,79% | 83,28% | 84,81% | 82,28% | 84,75% | 82,30% | 84,32% | 80,95% | 84,52% | 82,51% | 85,04% |
| Bayesian Boosting with Naive Bayes | 81,48% | 81,16% | 81,70% | 82,47% | 82,27% | 81,72% | 82,08% | 80,20% | 81,71% | 79,28% | 81,85% | 80,97% |
| AdaBoost with Naive Bayes | 82,23% | 80,69% | 81,25% | 82,20% | 82,62% | 81,58% | 81,94% | 80,06% | 81,73% | 79,60% | 81,95% | 80,83% |
| Bagging with Naive Bayes | 80,17% | 81,43% | 80,24% | 82,31% | 81,89% | 81,05% | 80,78% | 80,30% | 80,26% | 79,38% | 80,67% | 80,90% |
| Bayesian Boosting with K-NN | 85,74% | 82,94% | 86,36% | 84,11% | 86,44% | 82,27% | 87,20% | 84,56% | 86,74% | 84,20% | 86,50% | 83,61% |
| AdaBoost with K-NN | 85,74% | 82,94% | 86,36% | 84,11% | 86,44% | 82,27% | 87,20% | 84,56% | 86,74% | 84,20% | 86,50% | 83,61% |
| Bagging with K-NN | 85,74% | 83,12% | 86,42% | 83,47% | 86,69% | 81,99% | 86,89% | 84,93% | 87,30% | 84,52% | 86,61% | 83,61% |

Table C.1.40.: Percentage of wrongly classified cases of undersampled models, for rectal cancer.

| Wrongly Classified Cases - Rectal Cancer Undersampled Models | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1Year | | 2Year | | 3Year | | 4Year | | 5Year | | Average | |
| | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes | 18 attributes | 6 attributes |
| Stacking | 36,12% | 24,10% | 29,02% | 34,63% | 36,51% | 67,86% | 31,90% | 73,32% | 36,51% | 72,23% | 34,01% | 54,43% |
| Voting | 32,61% | 13,49% | 20,98% | 24,41% | 33,39% | 31,98% | 31,05% | 29,10% | 39,16% | 31,59% | 31,44% | 26,12% |
| Bayesian Boosting with Decision Tree | 67,71% | 17,24% | 32,68% | 31,12% | 67,32% | 33,15% | 68,33% | 37,52% | 63,57% | 37,60% | 59,92% | 31,33% |
| AdaBoost with Decision Tree | 67,71% | 17,24% | 32,68% | 31,12% | 67,32% | 33,15% | 68,33% | 37,52% | 63,57% | 37,60% | 59,92% | 31,33% |
| Bagging with Decision Tree | 31,59% | 12,40% | 24,57% | 27,93% | 67,39% | 33,93% | 71,68% | 29,02% | 65,76% | 32,84% | 52,20% | 27,22% |
| Bayesian Boosting with Random Forest | 35,49% | 13,96% | 29,49% | 26,99% | 46,88% | 37,75% | 48,28% | 30,81% | 51,25% | 29,17% | 42,28% | 27,74% |
| AdaBoost with Random Forest | 21,29% | 14,66% | 34,17% | 28,32% | 30,19% | 39,78% | 42,28% | 26,76% | 47,58% | 23,79% | 35,10% | 26,66% |
| Bagging with Random Forest | 36,97% | 12,56% | 33,62% | 26,83% | 50,70% | 33,46% | 43,29% | 25,66% | 39,47% | 29,95% | 40,81% | 25,69% |
| Bayesian Boosting with Naive Bayes | 75,98% | 34,87% | 57,88% | 35,26% | 76,68% | 39,47% | 74,34% | 42,28% | 72,78% | 43,68% | 59,61% | 32,59% |
| AdaBoost with Naive Bayes | 67,86% | 33,15% | 55,85% | 35,26% | 76,68% | 34,71% | 74,34% | 40,48% | 72,78% | 41,73% | 69,50% | 37,07% |
| Bagging with Naive Bayes | 48,60% | 27,07% | 48,05% | 35,18% | 76,60% | 39,24% | 74,34% | 40,33% | 72,78% | 42,67% | 64,07% | 36,90% |
| Bayesian Boosting with K-NN | 18,80% | 21,76% | 19,81% | 26,13% | 19,66% | 30,58% | 19,58% | 28,32% | 18,56% | 30,03% | 19,28% | 27,36% |
| AdaBoost with K-NN | 18,80% | 21,76% | 19,81% | 26,13% | 19,66% | 30,58% | 19,58% | 28,32% | 18,56% | 30,03% | 19,28% | 27,36% |
| Bagging with K-NN | 18,80% | 21,76% | 19,81% | 26,13% | 19,66% | 30,58% | 19,58% | 28,32% | 18,56% | 30,03% | 19,28% | 27,36% |