University of Brasília

Institute of Exact Sciences
Department of Computer Science

Explorando características relevantes do câncer colorretal com base em dados clínicos e biológicos: uma abordagem de bioinformática

Lucas Maciel Vieira

Thesis presented in partial fulfillment of the requirements for the degree of Doctor of Science in Informatics

Advisor
Profa. Dra. Maria Emilia M. T. Walter

Brazil

University of Brasília

Institute of Exact Sciences
Department of Computer Science

Explorando características relevantes do câncer colorretal com base em dados clínicos e biológicos: uma abordagem de bioinformática

Lucas Maciel Vieira

Thesis presented in partial fulfillment of the requirements for the degree of Doctor in Informatics

Profa. Dra. Maria Emilia M. T. Walter (Advisor)
CIC/UnB

Prof. Dr. Peter F. Stadler
University of Leipzig

Prof. Dr. João Carlos Setubal
University of São Paulo/São Paulo

Profa. Dra. Célia G. Ralha
University of Brasilia

Prof. Dr. André C. P. L. F. Carvalho
University of São Paulo/São Carlos

Prof. Dr. Ricardo P. Jacobi
Coordinator of the Graduate Program in Informatics

Brasília, 28th of February of 2023

To my parents,
Claudenir Mendes Vieira and
Hévila Maciel Mendes Vieira.
To my cousin,
Victor Hugo Vieira.

Acknowledgements

I would like to express my gratitude to my advisor, Prof. Maria Emília Machado Telles Walter, who helped me with patience and dedication, always available to share all her knowledge. For all the opportunities she gave me and for introducing me to the path of science, making me reach higher heights than I have ever imagined. For becoming one of my biggest professional references.

To Dr. Natasha, for all her support in the development of the project and for the period we worked together in Germany. To Prof. João Batista, for his support in the development of the work. To all those who participated and helped me, in some way, to carry out this work. I would also like to express my gratitude to all co-authors of the published article: Prof. João B de Sousa, Prof. João C. Setubal, Dr. Natasha A. N. Jorge, Prof. Maria Emilia M. T. Walter, and Prof. Peter F. Stadler.

To the University of Brasilia, essential in my professional training process and for supporting the preparation of this work.

"Biology is the study of complicated
things that have the appearance of
having been designed with a purpose."
Richard Dawkins

Extended Abstract

Colorectal cancer (CRC) is one of the most common and lethal types of cancer worldwide, being the second most common cancer in Brazil [1]. CRC is a heterogenous cancer that affects the lower part of the large bowel and can be classified according to its anatomical site as: colon, rectum, or rectosigmoid junction cancer. The most common type of CRC is adenocarcinoma, accounting for 90% of cases. Most CRC deaths are related to metastases and, if early detected, patient survival rates increase considerably. This disease can be impacted by many environmental factors, such as: eating habits, age, and weight. Treatment can differ depending on anatomical site and usually consists of surgery followed by chemotherapy. Inaccurate identification of the CRC anatomical site can lead to under or overtreatment, which can impact the patient's likelihood of mortality. The understanding of the molecular mechanisms and external factors that affect CRC development and progression is crucial to improving CRC prognosis, prevention, and treatment.

Considering the biological aspects of CRC, three types of coding and non-coding RNAs are of particular impact on the disease's underlying mechanisms. Highlighting: long non-coding RNAs (lncRNAs), micro RNAs (miRNAs), and messenger RNAs (mRNAs). In eukaryotes, mature mRNAs are formed after the pre-mRNA generated from the transcription undergoes a process known as splicing, which removes some regions (introns) of the pre-mRNA, while binding others (exons), thus forming the mature mRNA. The splicing process can generate more than one protein from a single gene in a process known as alternative splicing. The generated proteins are then used to regulate the organism's functions through use in metabolic reactions, by affecting many biological processes, such as disease development.

MiRNAs play an essential role in gene expression, by binding to mRNAs and initiating the inhibition or degradation of their target. In contrast, lncRNAs are not directly portrayed in this mRNA expression regulation process but play essential roles, such as altering other molecules' functions and therefore affecting protein expression and the development and suppression of disease. Given the specific role of each RNA described above in disease development, recent studies also highlight the importance of a mechanism known as competing endogenous RNA (ceRNA) networks, in which lncRNAs, miRNAs, and mRNAs

interact. In this mechanism, in addition to binding to mRNAs, miRNAs can also bind to ceRNAs, which then act as modulators of miRNAs and therefore indirectly regulate mRNA expression. The identification of ceRNA networks related to CRC development and its underlying mechanisms can help doctors better understand the disease and patient' prognosis. Some studies have been carried out using bioinformatic approaches to analyze and create ceRNA networks and to indicate potential prognosis biomarkers for colon, rectal, and CRC in general [2, 3, 4, 5, 6, 7, 8].

Although some studies were done with ceRNA network constructions in mind, to the best of my knowledge, this study is the first to establish specific ceRNA networks for: (i) colon; (ii) rectum; and (iii) rectosigmoid junction, and to relate them with specific biological mechanisms in order to identify differences and common factors between these sites.

Other studies suggest the use of machine learning methods using clinical features to predict CRC patient prognosis [9, 10, 11]. Specifically, Gründner et al. [9] explored a method that combines biological and clinical features to predict prognosis characteristics for CRC patients from South Africa. These studies showed promising results in predicting CRC patient' prognosis, but to the best of my knowledge, this study is to use open data and machine learning to predict CRC recurrence and patient survival by using biological markers extracted from the colon, rectal and rectosigmoid cancer ceRNA networks in combination with clinical features.

In this thesis, I begin by proposing a pipeline using open-access data from patients with CRC extracted from The Cancer Genome Atlas (TCGA) to construct CRC-specific ceRNA networks and potential biological markers that affect patient prognosis. Through analysis, I aim to identify RNAs that can be used as biological markers for the three CRC anatomical sites: colon, rectum, and rectosigmoid junction. To construct these networks and propose the biological markers, I collected RNA raw expression and clinical data from CRC patients. Using bioinformatic analysis tools to assess RNA expression profiles and building a ceRNA network for each CRC anatomical site, generated output in the form of ceRNA networks and the RNAs present on them. Next, through a functional enrichment analysis I assessed the potential biological pathways activated by the molecules obtained in the previous step. Finally, an overall survival analysis to identify the impact of these RNAs on patient prognosis, produced a list of potential biological markers as output.

Overall, the first pipeline of this thesis resulted in: the identification of several potential prognostic markers for colon, rectum, and rectosigmoid junction cancer; the construction of specific ceRNA networks for each anatomical site; and the identification of biological pathways that highlight differences in CRC behavior at distinct anatomical sites, thus reinforcing the importance of correct identification of tumor site. The output of this

pipeline consisted in a group of potential biological markers involved in CRC prognosis namely, the following site-specific prognosis biomarkers are of note: hsa-miR-1271-5p, *NRG1*, hsa-miR-130a-3p, *SNHG16*, and hsa-miR-495-3p in the colon; *E2F8* in the rectum; and *DMD* and hsa-miR-130b-3p in the rectosigmoid junction.

After generating the list of potential biological markers related to CRC prognosis, I proceed to the second part of this thesis: the proposal of a pipeline to predict CRC recurrence and patient survival using supervised machine learning (ML) methods. Clinical factors such as age and weight, as well as biological factors, can affect CRC progression and prognosis. To better CRC mechanisms and to identify the impact of both clinical and biological factors in prognosis, I used patient clinical features combined with the previously found biological markers as biological features to train the ML models. To improve predictive performance and interpretability of the proposed findings I evaluated and compared the following ML algorithms: Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), and Adaptative Boosting (AB). To establish the importance of each feature while building the models to predict CRC recurrence and patient survival, first, I performed a feature extraction analysis to filter and rank the actual impact of these features on the constructed prediction model. With the selected relevant biological and clinical features in hand, I then constructed the ML models and evaluated their performance. As output, this pipeline generated the ML models to predict CRC recurrence and patient survival along with a list of potential biological and clinical features relevant to patient prognosis.

Overall, the second pipeline resulted in the identification of several potential biological and clinical markers as important in CRC recurrence and patient survival. For feature importance: SNHG16, hsa-miR-130b-3p, hsa-miR-495-3p, and KCNQ1OT1 stood out as biological features; and age, ethnicity, pathological stage, chemotherapy, height and weight, positive lymph node count and lymph node count as clinical features. Finally, LR and RF achieved a best accuracy of 90% and 82% for predicting patient survival and CRC recurrence respectively. Also, the six proposed ML algorithms showed good performance overall, specifically, LR and RF displayed good overall results, which is coherent with findings from other studies [9, 10, 11].

This study strongly suggests that the use of bioinformatic approaches should be concurrently used with ML algorithms to enhance interpretation of CRC mechanisms and patient prognosis. However, some limiting factors are noteworthy: the amount of available data, being that the number of available patients for certain anatomical sites was low; and that the data mainly consisted of patients from the USA. Following the proposed pipelines, doctors can better understand the underlying mechanisms of CRC at its anatomical sites, and also use our model to help predict patient prognosis. Finally,

running these pipelines with Brazilian patient data could improve CRC data interpretation, especially in circumstances of diversity and inequality in a country's demographic landscape, which can affect CRC prognosis.

Keywords: non-coding RNAs, long non-coding RNAs, miRNAs, mRNAs, ceRNAs, machine learning, cancer, colorectal cancer

Resumo Expandido

O câncer colorretal (CRC) é um dos tipos de câncer mais comuns e letais em todo o mundo, sendo o segundo câncer mais comum no Brasil [1]. O CRC é um câncer heterogêneo, que se instala na parte inferior do intestino grosso e pode ser classificado de acordo com seu campo anatômico, como câncer de cólon, de reto ou na junção retossigmoide. O tipo mais frequente de CRC é o adenocarcinoma, que corresponde a 90% dos casos. A maioria das mortes causadas por CRC acontece quando esse entra em estado de metástase. No entanto, se detectado em seus estágios iniciais, a sobrevida do paciente com CRC pode melhorar consideravelmente. Esta doença pode ser influenciada por diversos aspectos ambientais, tais como: hábitos alimentares, idade e peso. Normalmente, o tratamento recomendado para pacientes com CRC é a cirurgia para sua remoção e, depois, o uso de quimioterapia, porém o tratamento pode diferir de acordo com seu campo anatômico. O diagnóstico do CRC em um campo anatômico incorreto pode levar o médico a prescrever um tratamento não recomendado ao paciente, o que pode afetar a sua taxa de mortalidade. Para auxiliar o prognóstico, prevenção e tratamento de CRC, é fundamental entender os mecanismos moleculares e os indicadores clínicos que afetam o desenvolvimento do CRC.

Quanto aos aspectos biológicos do CRC, podemos descrever o impacto dos RNAs codificadores e não-codificadores nos mecanismos subjacentes à doença. Em específico, podemos destacar três moléculas: RNAs longos não codificadores (em inglês, *long non-coding RNAs - lncRNAs*), micro RNAs (*miRNAs*) e RNAs mensageiros (em inglês, *messenger RNAs - mRNAs*). Nos eucariotos, os mRNAs maduros são formados a partir do pré-mRNA que, por sua vez, é produzido a partir do processo de transcrição passar por um processo conhecido como excisão (em inglês, *splicing*), que remove algumas regiões (íntrons) do pré-mRNA e liga outras regiões (exons), formando assim o *mRNA* maduro. O processo de *splicing* possibilita gerar mais de uma proteína a partir de um único gene, em um processo conhecido como excisão alternativa (em inglês, *alternative splicing*. Por sua vez, as proteínas coordenam quase todos os processos vitais no organismo, sendo utilizadas em reações metabólicas e afetando diversos processos biológicos, como o desenvolvimento de doenças.

Os *miRNAs* desempenham um papel essencial na expressão gênica, mais especificamente, ligando-se aos *mRNAs* e iniciando os processos de inibição ou degradação de seu alvo. Por sua vez, os *lncRNAs* não estão diretamente presentes neste processo de regulação da expressão de mRNA, mas desempenham papéis essenciais no organismo, como a alteração das funções de outras moléculas e, assim, afetam a expressão de proteínas indiretamente, o que pode contribuir para o surgimento e supressão de doenças. Considerando o papel específico de cada uma das moléculas descritas no desenvolvimento de doenças, estudos recentes destacaram a importância de um mecanismo conhecido como redes de RNAs endógenos concorrentes (em inglês, *competing endogenous RNAs - ceRNAs*), nos quais os *lncRNAs*, os *miRNAs* e os *mRNAs* interagem entre si. Nesse mecanismo, os *miRNAs*, que se ligam aos *mRNAs* pelos *binding sites*, podem também se ligar aos *ceRNAs*, assim, regulando indiretamente a expressão dos *mRNAs*. A identificação de redes *ceRNA* relacionadas ao surgimento do CRC e seus mecanismos subjacentes podem auxiliar os médicos a entender melhor a doença e realizar um melhor prognóstico do paciente. Na literatura, podemos encontrar alguns estudos que usam abordagens baseadas em bioinformática para criar redes *ceRNAs* e auxiliar a identificação de biomarcadores para o câncer de cólon, reto e o câncer colorretal em geral.

Embora alguns estudos tenham foco na construção de redes *ceRNA*, até onde sabemos, nosso estudo foi o primeiro a estabelecer redes *ceRNAs* específicas para: (i) cólon; (ii) reto; e (iii) junção retossigmóide, além de relacioná-los com mecanismos biológicos específicos, a fim de esclarecer as diferenças e fatores comuns entre essas diferentes localizações anatômicas.

Por outro lado, alguns estudos sugerem o uso de métodos de aprendizagem de máquina e também o uso de características clínicas para predizer marcadores que podem ser usados para prognóstico de pacientes com CRC [9, 10, 11]. Especificamente, Gründner et al. [9] sugeriram um método que combina características biológicas e clínicas para predizer marcadores de prognóstico de pacientes com CRC na África do Sul. Esses estudos descreveram bons resultados obtidos a partir de modelos de predição. Tanto quanto saibamos, nosso estudo foi o primeiro que usou dados abertos e métodos de aprendizagem de máquina para predizer a reincidência de CRC e a sobrevivência do paciente usando marcadores biológicos extraídos de redes *ceRNAs* de câncer de cólon, de reto e na junção retossigmoide, combinados com características clínicas.

Nesta tese, na primeira etapa, propusemos um *pipeline* utilizando dados de livre acesso de pacientes com CRC, extraídos do banco de dados *The Cancer Genome Atlas* (*TCGA*), para construir redes *ceRNAs* específicas para o CRC e marcadores biológicos que afetam o prognóstico do paciente. Nosso objetivo foi o de realizar uma análise para identificar moléculas que possam ser usadas como marcadores biológicos para os três sítios anatômi-

cos do CRC, cólon, reto e junção retossigmoide. Para construir tais redes e propor os marcadores biológicos, a expressão de RNA e os dados clínicos dos pacientes com CRC foram coletados. Os perfis de expressão de RNA foram produzidos por meio de ferramentas de análise que utilizam técnicas de bioinformática. Em seguida, encontramos redes *ceRNA* específicas para cada campo anatômico, para as quais, como dados de saída, obtivemos as redes *ceRNA* e as moléculas nelas presentes. Após essa etapa, foi realizada uma análise funcional, onde identificamos potenciais vias metabólicas relacionadas ao surgimento de câncer, as quais têm participação das moléculas obtidas na etapa anterior. Finalmente, uma análise de sobrevida global para identificar o impacto dessas moléculas no prognóstico do paciente foi realizada, resultando em uma lista de potenciais marcadores biológicos.

Nessa etapa, ficaram evidenciados diversos potenciais biomarcadores que afetam o prognóstico do paciente em câncer de cólon, de reto e na junção retossigmoide. Além disso, redes *ceRNA* específicas para cada campo anatômico foram construídas, e foram identificadas diferentes vias biológicas que destacam diferenças no comportamento do CRC nos diferentes campos anatômicos, reforçando assim, a importância de identificar corretamente o campo anatômico em que o tumor ocorre. Como resultados, geramos um grupo de potenciais biomarcadores biológicos que afetam o prognóstico do CRC, em particular, podemos destacar: hsa-miR-1271-5p, *NRG1*, hsa-miR-130a-3p, *SNHG16* e hsa-miR-495-3p para câncer de cólon; *E2F8* para câncer retal; e *DMD* e hsa-miR-130b-3p para câncer na junção retossigmoide.

Com a lista de potenciais marcadores biológicos que podem afetar no prognóstico de CRC, prosseguimos para a segunda etapa desta tese, em que propusemos um *pipeline* para prever a reinicindiva do CRC e a sobrevida dos pacientes, utilizando métodos de aprendizagem de máquina supervisionados. Fatores clínicos, como idade e peso, assim como fatores biológicos, podem afetar o prognóstico e o surgimento do CRC. Para melhor entender os mecanismos do CRC e identificar o impacto, tanto dos fatores clínicos, quanto dos fatores biológicos em seu prognóstico, usamos as características clínicas do paciente combinadas com os marcadores biológicos encontrados no passo anterior, como características biológicas, para treinar nossos modelos. Para alcançar um maior desempenho na predição e na possibilidade de interpretação dos resultados propostos, avaliamos e comparamos os seguintes algoritmos de aprendizagem de máquina: *Random Forest - RF*, *Logistic Regression - LR*, *Support Vector Machine - SVM*, *K-Nearest Neighbors - KNN*, *Decision Tree - DT* e *Adaptative Boosting - AB*. Para encontrar a importância de cada característica durante a construção dos modelos de predição, primeiro foi realizada uma análise de seleção de características, para filtrar e classificar quais dessas características de fato tinham impacto no modelo de predição construído. Com essas características

biológicas e clínicas relevantes selecionadas, construímos os modelos de aprendizagem de máquina e avaliamos seu desempenho. Finalmente, como resultado, geramos modelos de aprendizagem de máquina para prever a reincidência do CRC e a sobrevivência do paciente, e uma lista de potenciais características biológicas e clínicas relevantes para o prognóstico do paciente.

Nesta etapa, identificamos diversos potenciais marcadores biológicos e clínicos como importantes na reincindiva do CRC e na sobrevida do paciente. Quanto à importância das características, identificamos: SNHG16, hsa-miR-130b-3p, hsa-miR-495-3p e KCNQ1OT1 como características biológicas; e idade, etnia, estágio patológico, quimioterapia, altura e peso, contagem positiva de linfonodos e contagem de linfonodos como características clínicas. Finalmente, usando *LR* e *RF*, alcançamos uma precisão de 90% e 82% para predição da sobrevivência do paciente e da reincidiva do CRC, respectivamente. Além disso, o uso dos seis algoritmos de apredizagem de máquina propostos mostrou um bom desempenho geral, em específico, o *RF* apresentou bons resultados, o que também foi destacado em outros estudos [9, 10, 11].

Por fim, a pesquisa desenvolvida neste tese mostrou que o uso de técnicas de bioinformática em conjunto com o uso de algoritmos de aprendizagem de máquina pode melhorar a interpretação dos mecanismos presentes no CRC. No entanto, devemos destacar alguns fatores limitantes com os quais nos deparamos, como a quantidade de dados disponíveis para pacientes com câncer de junção rectosigmoide e a especificidade regional dos dados clínicos dos pacientes, visto que o banco de dados utilizado continha informações principalmente de pacientes dos Estados Unidos. Perspectivas de uso dos métodos desenvolvidos nesta tese são, primeiro, os *pipelines* propostos poderiam fornecer aos médicos um entendimento melhor dos mecanismos subjacentes ao CRC em seus diferentes campos anatômicos. Além disso, nossos modelos poderiam ser usados para auxiliar na predição de prognóstico do paciente. Por fim, executar esses *pipelines* com dados de pacientes brasileiros poderia ajudar os médicos a entender melhor as características específicas no surgimento do CRC e prognóstico dos pacientes que vivem nas diferentes regiões do Brasil.

Palavras-chave: RNAs não-codificadores, miRNAs, mRNAs, ceRNAs, aprendizagem de máquina, câncer, câncer colorretal

Contents

List of Figures

List of Tables

List of Acronyms

AB Adaptative Boosting

AI Artificial Intelligence

ceRNA Competing endogenous RNA

CRC Colorectal cancer

DE Differential expression

DT Decision Tree

KNN K-Nearest Neighbors

LASSO Least Absolute Shrinkage and Selection Operator

lncRNA Long non-coding RNA

LR Logistic Regression

miRNA Micro RNA

ML Machine Learning

mRNA Messenger RNA

PC Protein Coding

RF Random Forest

RFE Recusive Feature Elimination

SHAP SHapley Additive exPlanations

SVM Support Vector Machine

Chapter 1

Introduction

Two types of nucleic acid are found in nature, DNA (deoxyribonucleic acid) and RNA (ribonucleic acid). DNA stores information to generate amino acids and RNA molecules. Among the RNAs, some are expressed in proteins while many others do not code for proteins, but participate in many important cell functions. This last group is known as non-coding RNAs (ncRNAs). The fact that ncRNAs play important cellular roles is well known, e.g, chemical reaction catalyzes, gene expression, and chromatin regulation [12, 13]. In the human genome, less than 2% of the RNAs are transcribed into proteins, while the remaining RNAs exhibit no protein-coding function [14, 15, 16].

Messenger RNA (mRNA) is one of the types of RNA molecules transcribed into proteins, which via the translating process, acts as a protein blueprint used by the cell's machinery to build the protein. Proteins are molecules made up of amino acids, which play diverse essential roles in organisms, such as accelerating chemical reactions, transporting nutrients, eliminating toxic waste, and building complex structures [17]. Several studies designed to discover the role of specific proteins in cancer are pertinent to this thesis. For example, Her-2 is a protein related to increased proliferation and cell malignancy while inhibiting cells' differentiation and apoptosis [18]. Additionally, in cancer, several proteins have been shown to interact with ncRNAs, which can regulate gene expression [19].

NcRNAs can be classified as small ncRNAs, which are small in size (under 200 nucleotides) with some other known characteristics; and long ncRNAs (lncRNAs), longer than 200 nucleotides, which have almost no capacity to synthesize proteins and are, the least known transcripts [20, 21]. LncRNAs were once considered to be evolutionary junk or transcriptional noise [22], however, several studies have revealed the diverse biological roles played by lncRNAs in a variety of organisms (e.g., chromatin regulation [23]; growth limitation of the placenta in mammals [24]; regulation of good embryonic development [25]; regulation of the developmental cycle in the maturation of bone marrow [26]; and tumor suppression [27]). Among the presented roles of lncRNAs, disease regulation, especially

in cancer development is noteworthy [28, 29]. One type of small ncRNAs, microRNAs (or miRNAs) are responsible for the regulation of mRNAs, but also act to regulate the expression of genes responsible for development mechanisms, metabolism, cell proliferation, differentiation, and apoptosis [30], which can influence cancer initiation and progression. Regarding lncRNAs, Fachel et al [29], Beckedorff et al [31] and Reis et al. [32] show the involvement of lncRNAs in renal cancer cells, their participation in processes related to cancer epigenetics, and their potential as tumor indicators in the diagnosis and prognosis of cancer respectively. Several studies aimed to identify lncRNAs related to cancer using biological and bioinformatics techniques [33, 34, 19, 35, 36].

Cancer is a complex disease and one of the biggest causes of death in the world [37]. The emergence of the disease may be related to many factors, including genetic and epigenetic changes [38]. Increased understanding of the molecular mechanisms that cause the changes that initiate cancer is one of the most important aspects of cancer research [39]. Predicting the factors related to the emerging mechanisms of cancer can help prevent its development and facilitate its identification. In addition, the understanding of cancer mechanisms in cells can help identify cancer factors, which can be used to prevent it. To date, the literature in molecular biology and bioinformatics indicates lncRNAs and some classes of small ncRNAs (e.g., miRNAs) as biomarkers to understand cancer emergence [40, 21, 41, 42]. In addition to these biomarkers, clinical markers, e.g., racial, ethnic, or geographical, have been shown to be important in understanding the underlying mechanisms and behavior of cancer emergence [43].

This thesis focuses on colorectal cancer (CRC), one of the most common and lethal cancers in the world [44]. CRC occurs in the digestive tract, specifically in the colon, rectum, and rectosigmoid junction. The behavior and treatment of CRC can differ according to its anatomical site. Although prognosis, prevention, and treatment have advanced, due to the growing number of people diagnosed with CRC each year, better understanding of mechanisms in CRC development and progression continues to be crucial [3, 44].

Given the importance of mRNAs, miRNAs, and lncRNAs in cancer, recent studies also show the importance of their underlying interaction system in cancer progression, the so-called: competing endogenous RNAs (ceRNAs) mechanism [45, 46, 47, 48]. These molecules form a ceRNA network, which can play an essential role in cancer development [40, 49]. Therefore, exploring miRNAs, lncRNAs, mRNAs, and the ceRNA networks formed by them, along with clinical factors, could lead to a better understanding of the underlying mechanisms of CRC. In this context, this thesis aims to predict patient survival and CRC recurrence, by highlighting biological and clinical markers to characterize CRC behavior (and help in patient prognosis), taking into account the different anatomical sites: colon, rectum, and rectosigmoid junction. In more detail, it is common knowledge

that interactions among proteins, miRNAs, and lncRNAs influence cancer, since they can regulate suppressive and oncogenic functions in various types of cancer [40]. As previously stated, understanding these mechanisms may help prevent tumor emergence and cancer development, as well as facilitate its identification. Although several biological studies presented protein and ncRNAs' relationship with cancer [50, 22, 51, 52, 53, 54], few focus on predicting patient cancer-related prognosis by using protein, miRNA and lncRNA markers and patient clinical characteristics through computational techniques [9, 10, 11], despite the fact that several databases present disease-related information [55, 56, 57, 58] and others present cancer specific information [59, 60, 61, 39, 62].

Other studies relate the importance of patient clinical aspects for cancer, e.g., impact of race, age, and demographics on CRC emergence behavio [63, 64], but few explore the importance of these clinical aspects in combination with biological aspects to predict CRC recurrence and patient survival through machine learning.

1.1 Problem

To date, the use of computational methods to predict the importance of biological and clinical markers in the prognosis of colon, rectum, and rectosigmoid junction cancer is an open research problem.

1.2 Goals

The main goal of this thesis is to propose methods using bioinformatic tools, feature extraction, and machine learning techniques, to analyze the importance of clinical and biological markers in CRC prognosis.

The specific goals are:

- To build a reliable data repository containing proteins, miRNAs, and lncRNAs related to CRC;
- To propose a method based on interactions of miRNAs, lncRNAs, and proteins to infer ceRNAs, in order to analyze the importance of biological markers for the colon, rectum, and rectosigmoid junction; and
- To propose a method, based on supervised machine learning techniques and bioinformatics tools, to analyze the importance of biological and clinical features to predict patient survival and CRC recurrence.

1.3 Thesis outline

In Chapter 2, I first discuss the object of research, CRC. Then, I describe interactions between lncRNAs, miRNA, mRNAs, and ceRNAs and their relationships with CRC. After defining some machine learning techniques used in this thesis, focusing particularly on methods for feature extraction, I describe the bioinformatics tools and databases used in the proposed pipelines.

In Chapter 3, I propose a method to analyze the importance of biological markers for colon, rectum, and rectosigmoid junction, based on interactions of miRNAs, lncRNAs, and proteins to infer ceRNAs.

In Chapter 4, based on supervised machine learning techniques and bioinformatics tools, I devise a method to analyze the importance of biological and clinical features to predict patient survival and CRC recurrence.

Finally, in Chapter 5, I conclude this thesis and present future work.

Chapter 2

Background

In this chapter, I present the definitions, methods, and data used in this thesis. First, in Section 2.1, I briefly describe important aspects related to the object of study of this thesis, CRC. In Section 2.2, I describe the central dogma of molecular biology; lncRNAs, mRNAs, miRNAs, and proteins; and, based on the interactions of these genes, the competing endogenous RNA mechanism, called ceRNAs, and how they affect CRC progression. Next, in Section 2.3, I present basic concepts on machine learning and its paradigms, specifically exploring feature selection methods and how they are used in bioinformatics. Finally, in Section 2.4, I discuss methods proposed in the literature to predict ceRNAs and markers related to CRC progression as well as present databases containing information regarding ncRNAs related to cancer.

2.1 Biological aspects of colorectal cancer

In this section, I first describe the biological aspects of cancer, then I present some details on CRC, relevant to this thesis.

2.1.1 Cancer disease

Humans, in their adult phase, are estimated to have about 10^{15} cells, many of which need to divide and differentiate in order to replace other cells in organs and tissues [65]. This cell proliferation is required, for example, for embryogenesis, growth, maintaining the proper function of several adult tissues, and tumorigenesis [66].

The so-called stem cells are capable of dividing themselves, and it is estimated that they perform around 10^{12} divisions per day [65]. This process of proliferation creates many cell generations, and increased cellular multiplicity can sometimes be detected and called neoplasia (new growth) [65]. The neoplasia can be classified as: benign, when in spite of

presenting a high rate of cellar growth, the cells are still similar to normal tissues, and, do not generally have the ability to spread to other tissues and organs; and malignant, when cells tend to multiply quickly and spread to other tissues and organs, and even with proper treatment, recurrences of the disease are common. Malignant neoplasia can also be called cancer.

Cancer is a complex disease, one of the biggest causes of death and one of the main public health problems in the world [67]. The emergence of the disease may be related to many factors, including genetic and epigenetic changes [38]. Because genes can modify the birth rate or the death rate of individual cells, researchers have suggested their regulation as causative of the carcinogenic process [65]. One of the most important aspects of cancer research is the understanding of the molecular mechanisms that cause the changes that lead to its emergence [68]. In addition to improving to understanding of mechanisms of cancer cells, the prediction of factors can help to prevent it. Thus, several studies, both in molecular biology and in bioinformatics relate that ncRNAs and their interactions with proteins partake in mechanisms involved in cancer [69].

As previously stated, cancer can be characterized as a malignant neoplasm, where cells tend to multiply rapidly and spread to other tissues and organs. Because cancer can manifest in diverse tissues and organs, it can be considered a "group" of diseases. Even though all these diseases are characterized by abnormal cell division, each one presents specific peculiarities.

The understanding of genes and factors that affect its mechanisms is essential to better understand cancer. Although similar genes have been found to regulate processes in different types of cancer, some features that are specific to each type may be crucial to improving understanding of the disease. Taking this into account, the focus of this thesis is to propose a more curated prediction model to identify interactions of some genes involved in a specific type of cancer, CRC.

2.1.2 CRC disease

CRC is one of the most frequent and lethal types of cancer in the world [44]. According to its incidence, this disease manifests in three forms: family, hereditary, and sporadic. CRC can be classified based on three main affected sites, the colon, rectum, and rectosigmoid junction (Figure 2.1). Together, they form the large bowel. The colon is the largest portion, the rectum is located at the end, and the rectosigmoid junction is the transition between the colon and rectum [70]. A tumor site is classified as present in the rectosigmoid junction when differentiation between the rectum and sigmoid colon is impossible.

CRC can be impacted by many environmental aspects. Its development can be affected by unfavorable aspects (e.g., smoking, alcohol consumption, obesity, eating red meat,

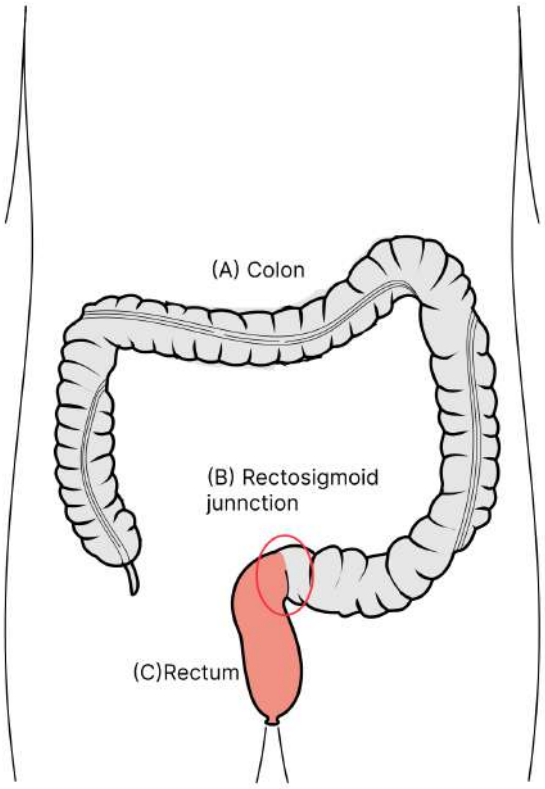

Figure 2.1 Three general main sites of lower bowel: colon (A), rectosigmoid junction (B), and rectum (C). The colon is the largest of the three anatomical regions, measuring around 180 cm, and can also be divided into: ascending, transverse, descending, and sigmoid colon. The rectosigmoid junction (B) corresponds to an anatomical site between the sigmoid colon and rectum measuring around 0.5 and 3.7 cm. The rectum (C) is the last part of the large intestine, measuring between 12 and 16 cm, and connects the colon to the anal canal. Adapted from [71].

diabetes, and sedentary lifestyle) and by favorable aspects (such as having a balanced diet including fruits and vegetables, whole grains, fibers, milk, and calcium). The impact of factors such as a balanced diet call attention to the patient demographic, as more developed populations tend to consume more processed food and consequently have a higher obesity rate, which can affect CRC development. Thus, development can also be impacted by other clinical factors such as race and gender. In this thesis, I also used some of these clinical factors associated with biological aspects.

There are a few groups of procedures that can help with CRC diagnosis: physical, endoscopic, and radiological examination. In the physical examination group, the doctor can perform anal inspection, rectal touch, and abdominal palpation, looking for signs of metastasis and abnormal behavior. Usually the development of CRC is difficult to

identify through physical examination. In the endoscopic examination group, the doctor can perform a rigid sigmoidoscopy, which explores up to twenty-five centimeters, does not require anesthesia or preparation, as well as colonoscopy, which is the gold standard for CRC diagnosis, but is expensive, requires sedation and is quite uncomfortable for the patient. In the radiological examination group, the doctor can perform: a virtual colonoscopy (which simulates a colonoscopy through tomography, but with less accuracy), tomography, and MRI (which detects lymph nodes). Although these exams are all used for CRC diagnosis, the one with most accuracy and that is most used is the colonoscopy, which can be expensive and uncomfortable for the patient, so it is usually not done unless it is truly necessary.

The most common type of CRC is adenocarcinoma, which accounts for 90% of cases. Most CRC deaths are related to metastases, and when CRC stays confined to the intestinal wall and is early detected it is potentially curable, as its development rate can be around five years [69]. A patient with CRC can be classified according to his or her individual risk factor by: medium, increased, and self-risk. Patients with medium risk are over 45 years old without other risk factors and are recommended to do a colonoscopy every five years and occult blood screening each year. Patients with increased risk either have a personal history of polyps or CRC, or have a first-degree relative with cancer or polyps and are recommended to do a colonoscopy and occult blood screening after 45 years old. Self-risk patients include members of families with polyposis, recommended to do a colonoscopy after puberty; members of families with lynch syndrome, recommended to do a Biennial colonoscopy after age 21 and annually from age 40 on; and people with Ulcerative colitis or Crohn's disease, recommended to do colonoscopies every one-two years, starting 8 years after diagnoses. As many countries do not have a prevention program, and the high cost of colonoscopy poses a challenge to frequent examination, other methods of CRC diagnosis in its early stages can improve treatment as well as the prediction of its course of development, the so-called, prognosis.

CRC can be classified through the TNM system (Table 2.1), where: T stands for tumor size; N stands for the spread to nearby lymph nodes; and M stands for metastasis to distant sites. The TNM system allows CRC to be classified into five pathological stages - 0, I, II, III, and IV. In stage 0, the patient has not been compromised and cancer cannot be considered invasive. In stage I the patient presents superficial tumors. In stage II the patient presents bigger tumors that already present an important penetration in the bowel wall, but do not present lymph node involvement. In stage 3 the tumor presents lymph node involvement. Finally, in stage IV the tumor presents distant metastasis. The lower the patient stage, the easier to treat the cancer. The patient stage can be assessed as a clinical-stage when based on pre-surgery information, and as a pathological stage

when based on post-surgery tumor information [72].

Table 2.1 TNM classification for CRC

Stage	**Level of involvements**
T2	Tumor involves muscularis propria
T3	Tumor beyond muscularis propria
N0	No involved nodes
N1	Up to three perirectal/colic nodes
N2	Four or more perirectal/colic nodes
M0	No distant metastasis
M1	Distant metastasis

CRC treatment is based largely on its stage and can differ depending on site. The recommended treatment for colon cancer is first surgery and then chemotherapy, most adjuvant treatment lasting around six months. Patients with rectum cancer usually undergo surgery and sometimes recieve chemotherapy and radiation before or after the surgery. In the case of the rectosigmoid junction, although it is anatomically considered part of the sigmoid colon, it shares the surgically important vascular system with the rectum, and is therefore better considered part of the rectum and treated independently [73]. Defining in which site the CRC occurs is important, given that inaccurate identification of the site can lead to undertreatment or overtreatment, which increases likelihood of mortality.

While CRC diagnosis and therapy have been progressing, given the growing number of CRC diagnosis in the population, improvement of prognosis is necessary [3, 74]. To better CRC diagnosis, prevention, and treatment, understanding the molecular mechanisms in CRC development and progression is crucial.

For this thesis, I collaborated with an expert on CRC who is a Professor at the Medical Faculty of the *Universidade de Brasília* (UnB), Professor João Batista. Through this collaboration, we chose features to propose the prediction model, the results of which have to be further validated experimentally.

2.2 Description of mRNAs, ncRNAs, and ceRNAs

In this section, I present biological concepts surrounding mRNAs and non-coding RNAs (ncRNAs). First, I briefly describe the biological aspects of messenger RNAs (mRNAs), small ncRNAs (in particular, miRNAs), and lncRNAs, and their role in cancer development. Then, I discuss the ceRNA mechanism and its potential roles in disease regulation.

2.2.1 The central dogma of molecular biology

Biological processes of regulation and structural maintenance occurring in organisms are driven by the interaction between two groups of molecules, proteins and nucleic acids. There are two types of nucleic acids in nature, DNA (deoxyribonucleic acid) and RNA (ribonucleic acid), which play essential roles in protein creation and system regulation.

As postulated by the central dogma of molecular biology [17, 75] (Figure 2.2), there are three main processes related to the interaction involving nucleic acids and proteins: (i) replication, in which a DNA strand is replicated, (ii) transcription, in which a portion of the DNA is transformed into RNA molecules, (iii) and translation, in which RNAs are used to produce a protein.

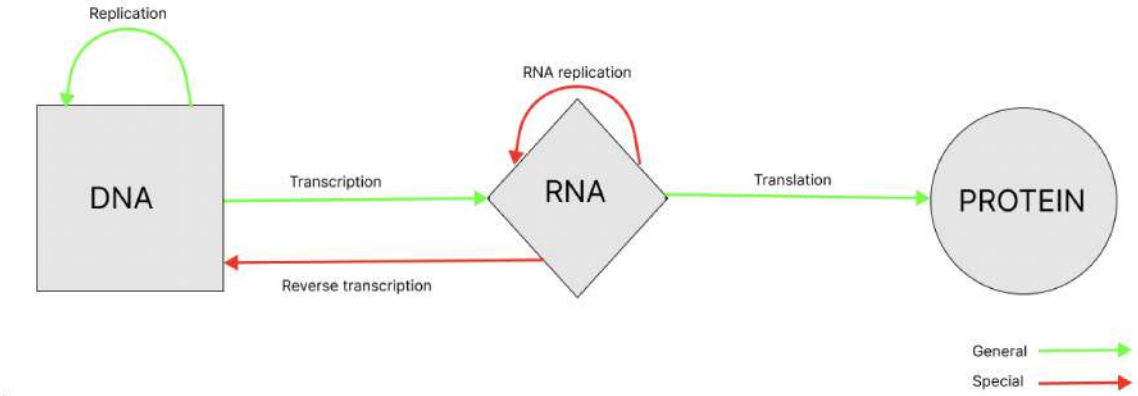

Figure 2.2 The central dogma of molecular biology explains the process of protein synthesis from information stored in DNA, performed with RNA molecules. This protein synthesis mechanism involves transcription, replication, and translation processes. In contrast to the central dogma flow, we have the reverse transcription and RNA replication processes.

The proteins generated after the translation process are molecules made up of amino acids, which play a variety of essential roles in organisms, accelerating chemical reactions, transporting nutrients, eliminating toxic waste, and building complex structures [17].

There exist many types of RNA molecules, which can play different roles in the cellular mechanisms [76]. RNAs can be divided into two groups: protein-coding (PC) RNAs, which can be translated into proteins, and non-coding RNAs (ncRNAs), which play regulation and structural roles in the cell.

Among the protein-coding RNAs, messenger RNAs (mRNAs) have been the major focus of research in biology for a long time. Among the ncRNAs, two categories are of note: (i) the small ncRNAs, which are small in size (20 to 300 nucleotides) and usually have known characteristics, (ii) and the long ncRNAs (lncRNAs), which are longer than 200 nucleotides, have almost no capacity to synthesize proteins, and are the least known

transcript [20, 21]. Next, I describe the biological aspects of mRNAs, small RNAs, and lncRNAs.

2.2.2 mRNAs

To better understand mRNAs, we must explore processes of the central dogma in more detail. During the replication process, the double-stranded DNA is separated into two strands by the helicase enzyme. When the separation of the strands begins, the transcription process is also initiated. In the transcription process, when the RNA polymerase identifies the promoter region, it guides the DNA transcription process in a messenger RNA that is not mature (pre-mRNA) in eukaryotes and in a messenger RNA (mRNA) in procaryotes [77]. In eukaryotes, the pre-mRNA generated by the transcription undergoes a process known as splicing. Splicing removes some regions (introns) of the pre-mRNA, while binding others (exons), thus forming the mature mRNA. Note that splicing can generate more than one protein from a single gene, known as alternative splicing (Figure 2.3). After the transcription process and splicing, translation begins, synthesizing a protein from the mature mRNA. The splicing process can play a key role point in numerous diseases, such as cancer [78].

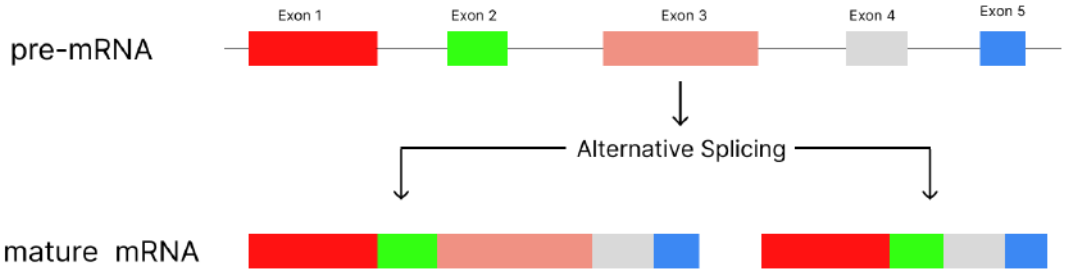

Figure 2.3 Process of alternative splicing, a mechanism that enables the production of multiple protein isoforms from a single pre-mRNA in which specific exons are included in or excluded from the mature mRNA [78].

2.2.3 Small ncRNAs

As cited, some specific classes of small ncRNAs are well known, and currently, there are almost 3,444 known classes (according to RFAM [79] in January 2023). The ncRNAs classes can also be classified according to their roles in an organism [80].

Small RNAs' functions are highly associated with their tertiary structure, which is derived from their secondary structure. When working with small RNAs, the secondary structure is used as an approximation [81, 79]. Some of best known classes of RNA stand out due to the roles they perform, such as: (i) tRNAs, which transport amino acids to assist in protein synthesis, (ii) rRNAs, responsible for the catalysis of protein synthesis, (iii) siRNAs, which can cause interference in protein translation, separating and promoting the degradation of stretches of mRNAs, (iii) snoRNAs, which can modify rRNAs, tRNAs, and snRNAs, (iv) and miRNAs, which regulates the translation process [80].

The functions of small ncRNAs in an organism include the development and suppression of diseases. For example, siRNAs can be used for the treatment of some diseases, such as cancer [82]. Also, miRNAs can regulate the expression of genes responsible for development, metabolism, cell proliferation, differentiation, and apoptosis mechanisms [83, 84], which can participate in the outset and progression of cancer.

Because they are used in the methods of this thesis, I will now discuss miRNAs in more detail. miRNAs are single-stranded small ncRNAs, normally approximately 22 nucleotides in length. As previously stated, miRNAs are known for regulating gene expression through translation inhibition or degradation of their target mRNAs in post-transcription [83] (Figure 2.4).

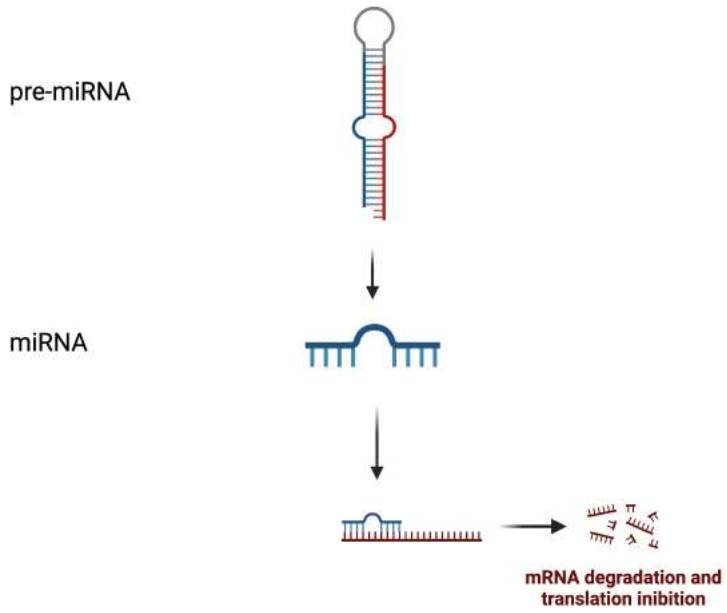

Figure 2.4 In the pre-miRNA process, the pre-miRNA is transformed into a miRNA, and regulates gene expression by binding to a mRNA and starting the process of inhibition or degradation of their target [85].

Processes like cell proliferation, differentiation, and apoptosis have also been proved to involve miRNA interactions [84]. Changes in miRNA expression can cause diseases, in particular tumor initiation, progression, and metastasis [86]. Several mechanisms, e.g., gene locus amplification, chromosomal deletion, mutation, and epigenetic silencing have been identified as responsible for deregulating miRNA expression in cancer [30].

2.2.4 LncRNAs

Unlike small ncRNAs, the lncRNAs are long molecules that are normally more than 200 nucleotides in length and have a poor capacity for synthesizing proteins. lncRNAs are usually classified into seven categories (Figure 2.5): (i) sense or (ii) antisense, when the lncRNA overlaps the transcription region of one or more exons of another gene, on the same or the opposite strand, respectively; (iii) bidirectional, when the start of the lncRNA transcription and another gene in the opposite strand are close; (iv) intronic, when the lncRNA is derived entirely from introns; (v) enhancer, when the lncRNA is located in enhancer regions[1]; (vii) intergenic, also called lincRNA, when the lncRNA is located between two genes; (vii) or promoter, when the lncRNAs are located in promoter regions [87][2].

Unlike the studies on small ncRNAs, which use their secondary structure as a starting point to infer their cellular roles, given the long length of lncRNA, research is limited by current tools to build the spatial models of these molecules. Even with the restriction posed by limited analysis of their secondary structure, several studies show that lncRNAs can interact with DNA, proteins, and other RNAs in transcription processes, therefore sharing responsiblility for growth, differentiation, suppression, and establishment of cells, which are usually deregulated in cancer [87, 89].

Yet, lncRNAs can exhibit an mRNA-like structure with a poly-A tail[3] in certain cases, and can exert roles to act as: (i) decoy, by binding to other RNAs and proteins to alter their functions; (ii) scaffold, by connecting chromatin-modified proteins and DNA regions to form signal connections; (iii) guide, by miRNA sequestration; and (iv) signal, by modulating miRNA regulation [89].

Like small ncRNAs, lncRNAs play diverse roles in an organism, which can include participation in the development and suppression of diseases. Thus, several studies, both in molecular biology and in bioinformatics propose lncRNAs as biomarkers to understandand cancer emmergence [69, 50, 22, 51, 52]. These studies show the role of lncRNAs in

[1]Enhancer region is a region of DNA that amplifies transcription by interplaying with their target promoters.

[2]Pomoter region is a region of DNA in which molecules bind to initiate transcription.

[3]A poly-A tail is a long chain of adenines, which are added to a RNA.

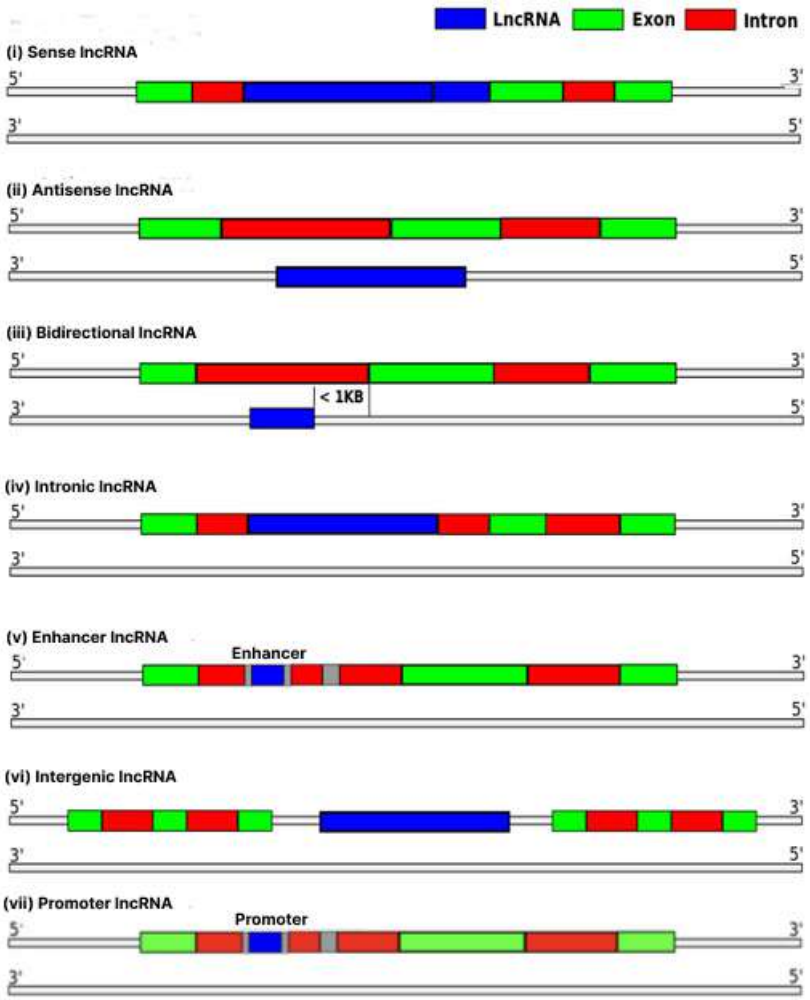

Figure 2.5 Long non-coding RNA classes: (i) sense; (ii) antisense; (iii) bidirectional; (iv) intronic; (v) enhancer; (vi) intergenic; and (vii) promoter. Adapted from [88].

epigenetics as well as the interaction between lncRNAs and miRNAs in cancer development.

2.2.5 Competing endogenous RNAs (ceRNAs)

As previously described, proteins play diverse essential roles in organisms, any disturbance in their expression could lead to different behavior in the organism, such as cancer. The mRNAs and the miRNAs are directly related, miRNAs bind to mRNAs and can affect the translation process, therefore, both can directly affect normal protein expression. Although lncRNAs are not directly involved in the described process, they can also interact with miRNAs and mRNAs, therefore indirectly affecting gene expression and also potentially contributing to cancer development. This lncRNA-miRNA-mRNA interaction is seen when, in addition to the regulation of mRNAs by miRNAs, the miRNAs bind to

lncRNAs. These interactions act as a new miRNA regulation mechanism, impacting the ability of other RNAs to compete for miRNA binding sites[4] [30, 90]. This mechanism is also known as the "target mimicry" process, while other RNAs, e.g., lncRNAs, act as miRNAs "decoys" ("sponges") and affect transcriptional regulation [30]. Simply put, the sponge shares a binding site target with a miRNA, and when the sponge interacts with the targets, it releases them from the miRNA, affecting miRNA mechanisms (Figure 2.6). This mechanism is also called competing endogenous RNAs (ceRNAs). Poliseno et al. [91] and Sumazin et al. [92] show this sponge behavior in cancer emergence.

For example, lncRNA's role as a decoy when interacting with miRNAs has been pointed as possibly responsible for the emergence of lung, prostate, breast, pancreatic, CRC, and other types of cancer [94, 95, 87, 89, 42]. Another example, for CRC, Zhang et al. [96] point to the interaction among the PC SIX4, the lncRNA H19, and miRNA miR-193b-3p. In this interaction, H19 acts as ceRNA for the miR-193b-3p, which affects SIX4 regulation and affects transcription regulation. Other studies, such as Lin et al. [97], Li et al. [98], and Gao et al. [3] also describe the roles played by lncRNA.

2.3 Concepts of machine learning and feature selection

In this section I present machine learning (ML) concepts, exploring feature selection definitions and methods. First, I present basic concepts and the learning paradigms of ML and detail ensemble learning and random forest, which are both used in this thesis. Then, I highlight the importance of data in ML, along with general techniques commonly used to build a prediction model.

2.3.1 Learning paradigms

ML focuses on the development of algorithms that detect patterns and learn through experience. According to Russel et al. [99], a machine learns when it improves its performance in future tasks from observations made in the past. Machine learning is a category of artificial intelligence algorithms through which, given an input, a machine becomes accurate in predicting outcomes, and then produces an output without having been explicitly programmed for these tasks. ML has four learning paradigms - unsupervised, supervised, learning by reinforcement, and semi-supervised, briefly described as follows.

Supervised algorithms are based on the knowledge of the classes being analyzed. Basically, it classifies input data into previously known classes. To do so, the method builds

[4] A binding site is a region on a macromolecule (e.g., protein) that binds to another molecule.

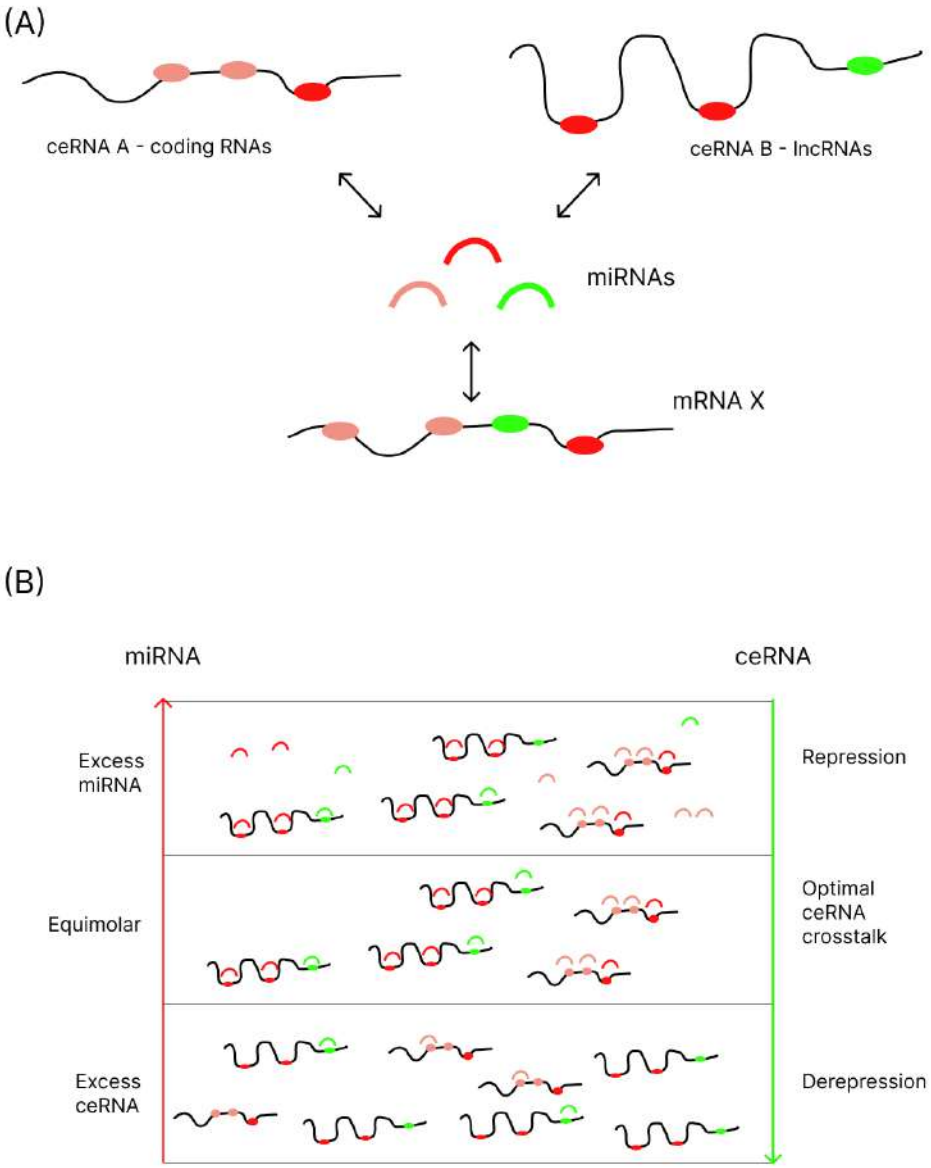

Figure 2.6 Competing endogenous RNAs mechanism. (A) illustrates the interaction among ceRNA molecules, where RNAs compete for miRNAs affecting the splicing and a different mRNA [90]. (B) describes the scenarios in which the miRNA and the ceRNAs are over and under-expressed, affecting the regulation of molecules in the organism. Adapted from [93].

a function called hypothesis that, according to the features, maps the input to the output. This function is used as a model that becomes capable of classifying new input data as belonging to specific classes. Examples include support vector machine (SVM) [100],

Decision tree (DT) [101] and K-nearest neighbors (KNN) [102].

Unsupervised learning recognizes patterns in an input collection. The classes of the input collection are not used in the algorithm process, even if they are known. Based on the input features, the algorithm searches for patterns in the input data, labeling portions of the data, which have been recognized as a class. The output is composed of groups of input data. Some examples are k-medoids [103] and k-means [104].

Learning by reinforcement is based on learning each interaction to achieve a final goal. The algorithm interacts with the environment, which is characterized by elements other than the program itself. A decision made by the program receives a score, used to decide the best classification. The decisions taken by the program receive rewards, which inform the best action to take, given the possible known states of the environment [105].Some examples are SARSA [106] and LSTD [107].

Lastly, the semi-supervised learning paradigm is method that extends supervised learning by using unsupervised learning techniques. In some cases, its performance exceed both unsupervised and supervised learning approaches used separately [99]. Some examples are Label Spreading [108] and Label Propagation [109]. Given the learning paradigms, I next describe the supervised ML algorithms adopted in this thesis.

2.3.2 Logistic Regression

Logistic Regression (LR) is a supervised ML algorithm based on the statistical technique with the same name. The LR algorithm is a special case of linear regression, which creates a function from the available features to map the input data to the targets to predict the probability of a new data point belonging to the target class [110]. In this thesis, I used LR because of its simple implementation and interpretation, and for its calculation time speed.

2.3.3 Support Vector Machine

Support Vector Machine (SVM) is a supervised algorithm that classifies groups based on the creation of separation margins. These margins, found by a fraction of the training data, are called support vectors, and they separate sets of data into known labeled classes (Figure 2.8). In this thesis, I used SVM based on its good capacity for generalization, reducing classification errors, and for working well with a low amount of data (when compared to other ML methods).

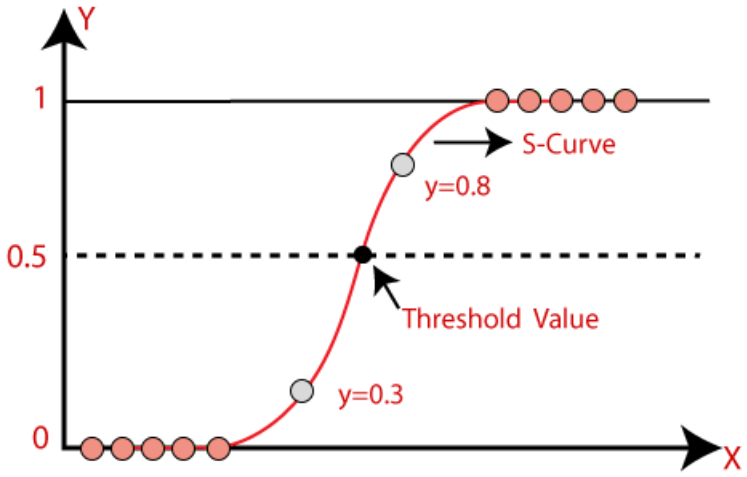

Figure 2.7 Example of logistic regression curve. The S-curve is generated from the logistic function, which estimates the probability of the target. The probabilities are bound between 0 and 1. Adapted from [111].

2.3.4 K-Nearest Neighbor

K-Nearest Neighbor (KNN) is a non-parametric lazy learning supervised algorithm, based on a parameter K, which represents the number of neighbors that influences the classification. The distance among the input data generates a classification model. KNN plots the input data in a feature space where we have a notion of distance.

Basically, KNN finds a group of K objects in the training set that is closest to the test input data object and labels each point as belonging to a particular class in this neighborhood. There are three major parts to this algorithm: a set of labeled input data; a distance metric to compute the distance between two data points; and the value of K, the number of nearest neighbors [112]. Figure 2.9 shows an example of KNN. In this thesis, I used KNN given its simple implementation and interpretation, and for its calculation time speed.

2.3.5 Decision Tree

Decision Trees (DT) [101] is a non-parametric regression tree estimator that embeds tree-structured regression models into a well-defined theory of conditional inference procedures. The DT structure is represented by two types of nodes: (i) the decision node, which is used to represent a decision based on a feature, and (ii) a leaf node, which is the decision output. The root node is the starting point, which further expands to various branches

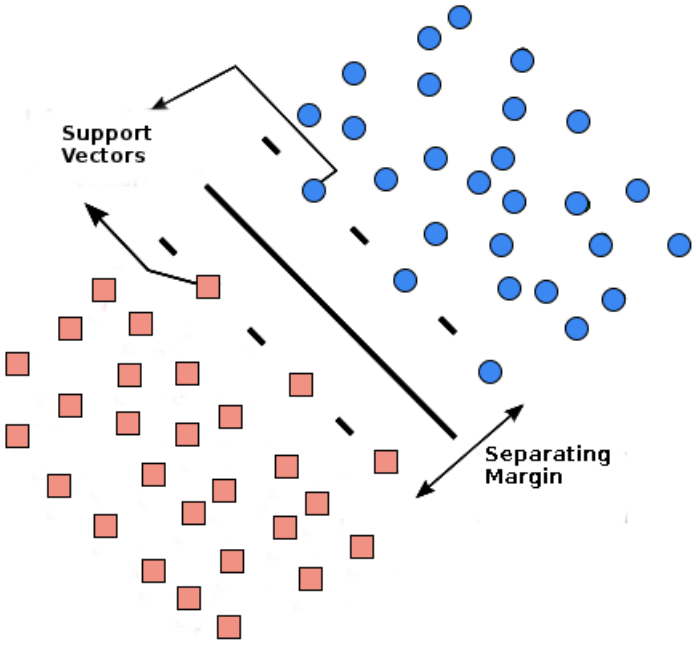

Figure 2.8 Example of support vectors with dimension 2, where the support vectors separate circles from square objects. Adapted from [88].

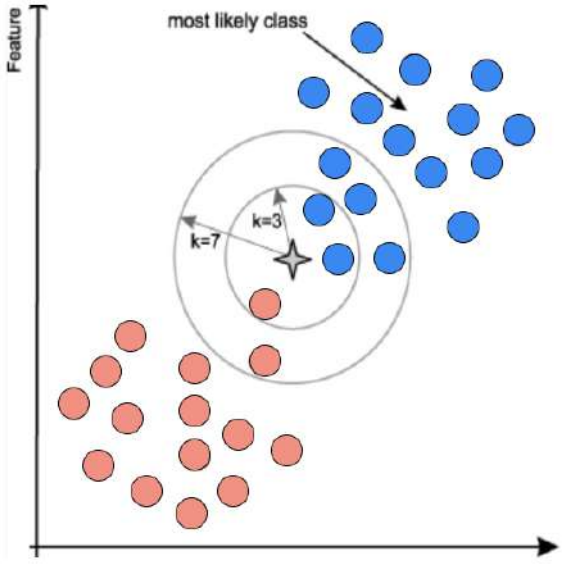

Figure 2.9 Example of KNN with neighbors influence example, for $K = 3$ and $K = 7$. Adapted from [113].

making a tree-like structure. On each decision node, the data point class is defined through hierarchy, based on a yes or no question [114]. I used DT in this thesis because of its simplicity in ranking and, interpretation, and for its calculation time speed.

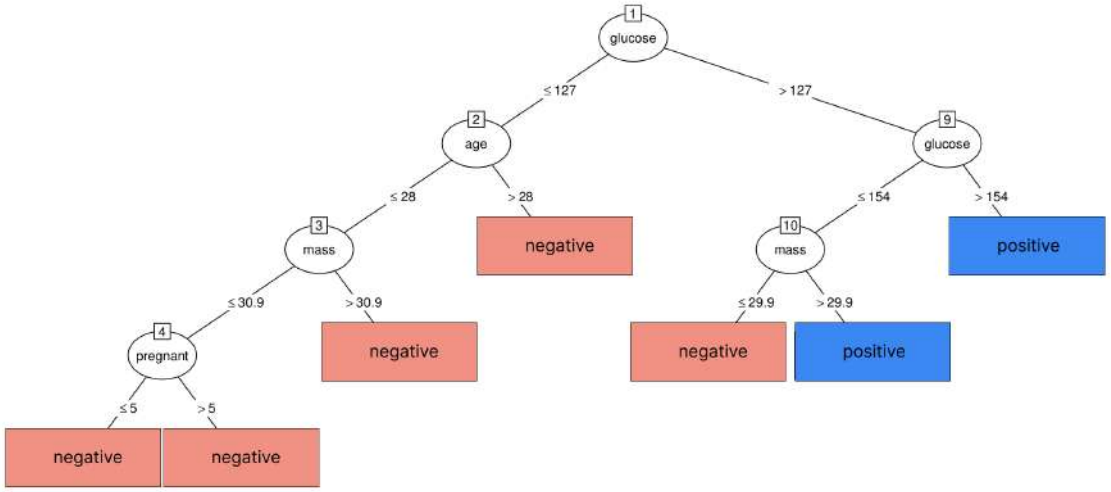

Figure 2.10 Example of DT. The higher the tree node the more relevant the feature in data classification. In this case, when input data has a glucose value less than or equal to 127, it is most likely labeled as negative. Adapted from [115].

2.3.6 Ensemble

Ensemble based systems consult many sources before making a decision, given their known variability and accuracy in other records [116]. This consulting happens because, most of the time, knowledge from a single source is not enough to make trustworthy predictions, and complementary knowledge from other sources can improve predictions.

Therefore, the ML method called Ensemble [117] uses a combination of a set of classifier estimators, to build a classification model with improved generalizability, as compared to a single estimator model. Ensemble methods are known for their capacity for generalization, which mostly achieves better results than when results are obtained through independent execution of each method.

Based on how the learners are generated, the ensemble methods are divided into two paradigms: sequential ensemble (boosting methods); and parallel ensemble (averaging methods) [117]. The averaging methods hinge on the construction of several parallel independent estimators, which produce output in the form of a prediction based on the average of their estimators. Normally, their combination is better than a single estimator because the model variance is reduced. On the other hand, in boosting methods, where several estimators are built sequentially, to reduce the bias of these estimators' combinations.

When using boosting, the combination of several weak models can produce an improved model, while averaging methods work better using the combination of strong esti-

mators. Ensemble learning is a well fabricated and frequently used ML method for classification and can perform well in fields such as feature selection [118]. Given the learning paradigms and metrics for the performance measurement, I next describe Random forest (RF) and Adaptative boosting (AB) which are the ensemble algorithms adopted in this thesis.

2.3.7 Random Forest

RF is thus named because it builds a number n of decision trees as an ensemble, to create a better classification model. A RF is an estimator that fits decision tree classifiers on various subsamples of the dataset, also using averaging to improve the predictive accuracy while simultaneously controlling overfitting [117][5].

In a random forest, each tree in the ensemble is built from a sample in the training set. In addition, when splitting a node, the chosen split is no longer the best split among all the features. Instead, the split that is chosen is the best one among a random part of the features [117]. As a result of this randomness, the bias of the forest usually increases slightly but, due to averaging, its variance also decreases, usually more than compensating for the increase in bias and yielding an overall better model. The feature importance can also be extracted by the analysis of the relative rank of a feature used as a decision node in a tree. Features used at the top of the tree are most significant in the final prediction. Figure 2.11 shows a RF example. I used RF in this thesis because of its simplicity in performing feature ranking and its reduced prediction error rate.

RF is a ML algorithm that can be used for feature selection because it can identify the features that are most important in predicting the target variable. This is done by training a large number of decision trees on different subsets of the data, and then averaging the predictions made by each tree (Figure 2.11).

By comparing the relative importance of each feature, measured by the number of times it is selected by the trees, the RF algorithm can rank the features according to their importance. This can be useful in identifying the most important features, which can then be used to build a more interpretable and efficient model. One of the advantages of using RF for feature selection is that it can handle high-dimensional data, such as transcriptomics data, which may have hundreds or thousands of features. It is also robust against noise and outliers in the data, making it suitable for use with real-world data that may contain errors or irregularities.

[5]Overfitting is when a model is accurate in predicting training data, but not capable of generalization for predicting new data.

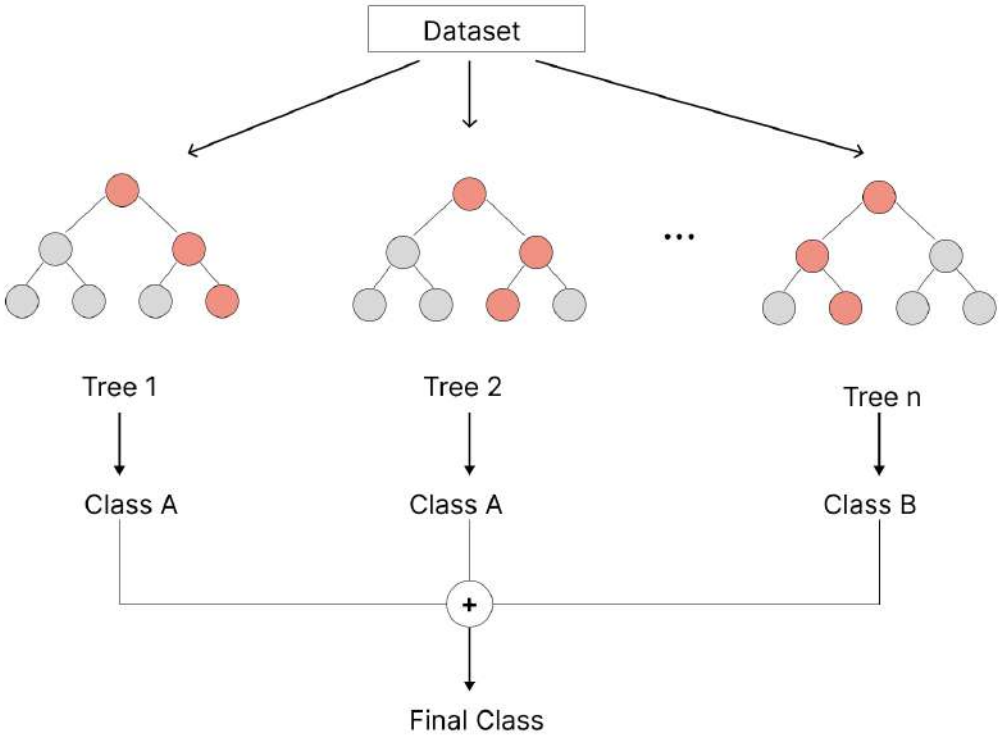

Figure 2.11 RF prediction model in which a large number of decision trees are generated from a dataset. Each decision tree prediction is combined to make a final prediction. Each decision tree is trained on a random subset of the data, and the final prediction is made by averaging the predictions of all the decision trees.

In summary, RF is a useful tool for feature selection in transcriptomics because it can identify the most important features in a high-dimensional dataset, and it is robust against noise and outliers in the data.

2.3.8 Adaptative Boosting

Adaptative Boosting (AB or AdaBoost) is a boosting algorithm that can be used to improve the accuracy of other algorithms. It works by repeatedly training the weak learner with different weights given to each training data point, where more weight is given to the examples that the weak learner previously misclassified (Figure 2.12). The final output of the AdaBoost algorithm is a combination of all the weak learners, where each weak learner votes on the outcome with a weight proportional to its accuracy [119].

In this thesis, I used AdaBoost combined with the DT estimator. The main difference between this approach and RF is that in AdaBoost the constructed trees are usually just a node and two leaves, meanwhile, in RF, a full-sized tree is built in each interaction. I

used AdaBoost in this thesis because of its ability to build a strong learner from multiple weak learners.

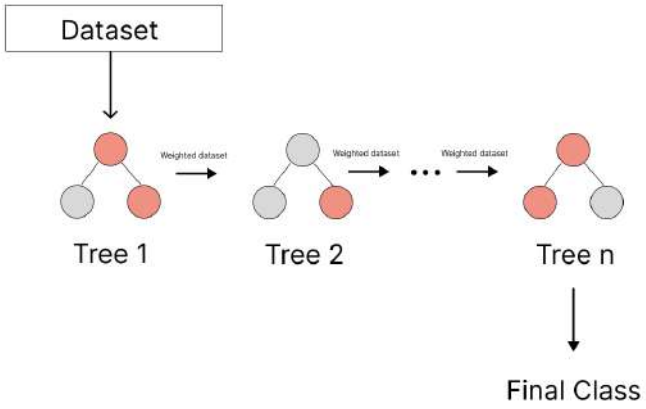

Figure 2.12 Example of the boosting technique used in AdaBoost, a prediction model in which a large number of decision trees are generated from a dataset. The first tree receives the dataset as input adjusts the model according to a feature, weights the dataset, and gives the adjusted knowledge to the next tree. The last adjusted tree makes the class prediction.

2.3.9 Feature selection and ranking

Feature selection is the process of selecting a subset of relevant characteristics to be further used in ML methods. This is often done in order to reduce the data dimensionality, improve interpretability of results, and increase analysis efficiency. Several different techniques can be used for feature selection, such as: filter methods; wrapper methods; embedded methods; and hybrid methods [120]. Next, I briefly describe these techniques.

Filter methods involve ranking the features based on some statistical measure of relevance, such as the $p-value$ from a $t-test$, and then selecting a subset of the top-ranked features. Wrapper methods evaluate the performance of a predictive model on a subset of the features and then select the subset that leads to the best performance. Embedded methods involve learning the features and the model jointly, such as using regularization to select features that are more important for prediction. Finally, the hybrid methods combine elements of multiple feature selection techniques, such as using a filter method to pre-select a set of features and then using a wrapper method to further refine the selection.

The fact that working with datasets containing a large number of features is common and it can present some challenges being that some ML methods do not deal well with high dimensionality. There are also cases in which the specialist helping to build the

model needs to understand the features (i.e., doctor, biologist, and others) to check for business value.

Feature selection is a field of itself, which focuses on helping produce a correct selection of features that can improve the inductive learner, either in terms of learning speed, generalization, or simplicity [121]. Feature selection is commonly divided in two ways: one that returns a subset of the giving features; and another that returns a ranking of the giving features.

One example of a feature selection method is Recursive Feature Elimination (RFE), which tries to select the optimal features for a model based on its accuracy [122]. RFE is often times, used in combination with custom regressors to optimize feature selection. Specifically, the Least Absolute Shrinkage and Selection Operator (LASSO) [123] is widely used. LASSO is a statistical formula that aims to identify the used variables and to assign coefficients to them, leading to a model with minimal prediction error. Given the implicit feature ranking capability of feature selection methods and of the previously described RF, we can further explore RF to better understand each feature's impact on a model's prediction, as well as use RF in combination with RFE.

2.3.10 SHAP

Contrary to the common view of ML models as a black box, Shapley Additive exPlanations (SHAP) is a mathematical method for explaining the output of ML models. SHAP estimates feature attribution for individual runs and captures the contribution of each feature given a prediction model [124].

SHAP is a mathematical method for explaining the output of a ML model based on the idea of Shapley values, which come from game theory and measure the contribution of each feature to the prediction made by the model. In simple terms, SHAP allows for an understanding of how each feature of a model contributes to the final prediction made by the model. For example, given a model that predicts whether or not a customer will get a loan, SHAP can help to understand which features of the customer's data (e.g., income and credit score, among others) are most important in making that prediction.

This method works by calculating the contribution of each feature to the final prediction made by the model by comparing the model prediction with and without each feature and then quantifying the difference in the prediction. This sheds light on how much each feature contributes to the overall model prediction. Therefore, SHAP can be used to explain the predictions made by any ML model, including DT, RF, and LR. Also, it is a powerful tool for improving the transparency and accountability of ML systems. SHAP estimates feature attribution on individual runs and captures the contribution of each feature given a prediction model [124].

2.3.11 Generic machine learning techniques

In this section, I describe the ML techniques used in this thesis.

Input data preparation

Here, I address the importance of having a good quality data collection. Gathering the data collection is considered the starting point of building a ML model. As previously stated, ML is a process that, given an input, allows a machine to predict outcomes and produce an output. If an input has mislabeled data, the ML function will probably map a new input to output incorrectly, since it learned from erroneous observations. Additionally, if we build a model with just one input, the model will be unlikely to learn enough to classify other data differently from this single input. These aspects lead to very relevant problems when building a ML model, and introduce the issues of data with noisy information and not having a suitable amount of data.

Techniques to remove outliers (or noisy information) should be used in building the input database in order to minimize this problem. These techniques can vary from simple approaches, like removing points given the standard deviation, to the use of unsupervised learning to cluster data and use only points near a centroid. Regarding the amount of data, there are several ways to minimize the problem of inappropriate amounts of data, using techniques that vary from under-sampling the data collection (e.g., random under-sampling) to over-sampling the data (e.g., SMOTE [125]). Aside from these techniques, when building the input data collection, the evaluation of a specialist is always recommended, for example, working with a biologist when building a model to predict RNAs.

Another problem that must be considered is the selection of suitable input data groups when working with supervised and semi-supervised learning techniques. For example, in classical binary (two classes) classification problems, data is normally divided into two classes, positive and negative. For example, if a model is built to classify a data point as a car, data containing cars are chosen as positive. Now, if the negative group is defined as fruit, the model can build a hypothesis that classifies a data point as a car with high accuracy, using features such as tires and windows. But, if the model receives a motorcycle as input, it will probably be erroneously classified as a car. This problem arises when the negative group is not wisely chosen. Basically, it classifies the input data as belonging to one of the previously learned classes.

Despite the recent increase in related biological data, because many biological database entries are not fully annotated, sparsity of available datasets continues to be a problem, which may not be ideal for statistical learning and construction of ML models [126]. Also, the privacy aspects inherent to human data create a further problem of limited access to

data. The fact that diseases occur only in a small fraction of a population further limits the portion of biological data. This can lead to imbalanced data, which can generate overfit models, and false-positive findings [127]. The participation of specialists in curating the data, giving biological meaning, and using other performance metrics can improve ML methods applied to biological data [126, 127].

Training, testing, and validation phases

Usually, there are three steps in the creation of ML models, training, validation, and testing. Training is the step that aims to generate the model [117], and through which the prediction hypothesis is built. This step uses a fraction of the input data, which is not used in the testing step. In simple terms, training is the step where the model learns.

In the validation step, the hyperparameters of the model created in the training step are tuned. Again, a fraction of the input data is used that is not used in the training and testing steps.

After building a model through the training step, testing is performed in order to validate the prediction hypothesis. Yet again, a fraction of the input data is used that is, not used in training and validation. In this step, the model's prediction performance is calculated.

Performance measures

In addition to classification, and assuring the good input data quality, predictions made by the model must be verified to be correct. Most metrics used to analyze the quality of the built model, are usually calculated based on the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), from the output classification of the constructed model in the testing phase. TP (TN) shows the outputs where a test observes a positive (negative), and a positive (negative) was also predicted. FP (FN) shows the outputs where a test observes a positive (negative) for a negative (positive) input, or the model predicted as a positive (negative), a negative (positive) input. These numbers are usually shown in the so-called confusion table (Table 2.2), often used to visualize the performance of the model.

Table 2.2 The confusion table, shows the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) predicted by the model constructed in the training phase.

	Predicted Value	
Real Value	Yes	No
Yes	TP	FN
No	FP	TN

The confusion matrix calculates metrics that evaluate the performance of the model constructed in the training phase. Some commonly used metrics are recall, precision, specificity, F-measure, and accuracy (each of which measures a particular aspect of the built model).

Recall shows the rate of positives predicted as so, and it is calculated by:

$$recall = \frac{TP}{TP + FN}$$

Precision shows the rate of input data classified as positive, which are really positive, and is calculated by:

$$precision = \frac{TP}{TP + FP}$$

Specificity calculates the rate of negatives predicted as so, and it is calculated by:

$$specificity = \frac{TN}{TN + FP}$$

F-measure combines precision and recall using a harmonic mean, and it is calculated by:

$$F - measure = \frac{2 \cdot Precision \cdot recall}{Precision + recall}$$

Finally, accuracy is a metric that calculates the general rate of the model:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

The performance can also be represented by the use of charts, in order to facilitate interpretation. For example, the Receiver Operating Characteristics (ROC) curve can help in understanding the ratio of true positives to false positives by analyzing the Area Under the Curve (AUC), where the closer results are to 1, the better the model.

As previously stated, model performance is an important aspect of data classification. In this thesis, I cite two techniques that can be used to improve performance of a built model: k-fold cross-validation and grid search. In k-fold cross-validation, data is divided into k segments (folds) of equal size. Then, k training and testing iterations are performed, and in each iteration, a segment of the data is used as validation, while the other $k - 1$ segments are used for training. During the process, the segments are rearranged to ensure that each segment is representative [128]. Figure 2.13 shows an example of the use of k-fold cross-validation, with k = 10.

The grid search method performs a hyperparameter optimization, in order to improve the model performance, by manually setting a range of possible parameters for the chosen

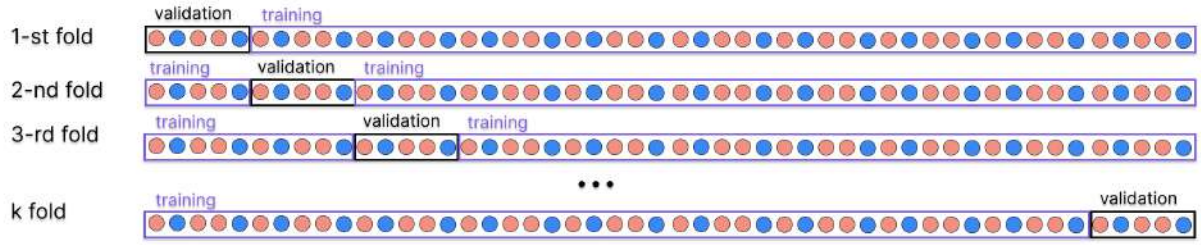

Figure 2.13 Cross-validation, illustrated on a data set containing k = 10 segments. Each segment in turn serves as a single validation segment. The model is built using the remaining k - 1 segments.

model algorithm and searching in a brute-force way for its optimal parameters according to the performance metric established [129]. See an example in Figure 2.14.

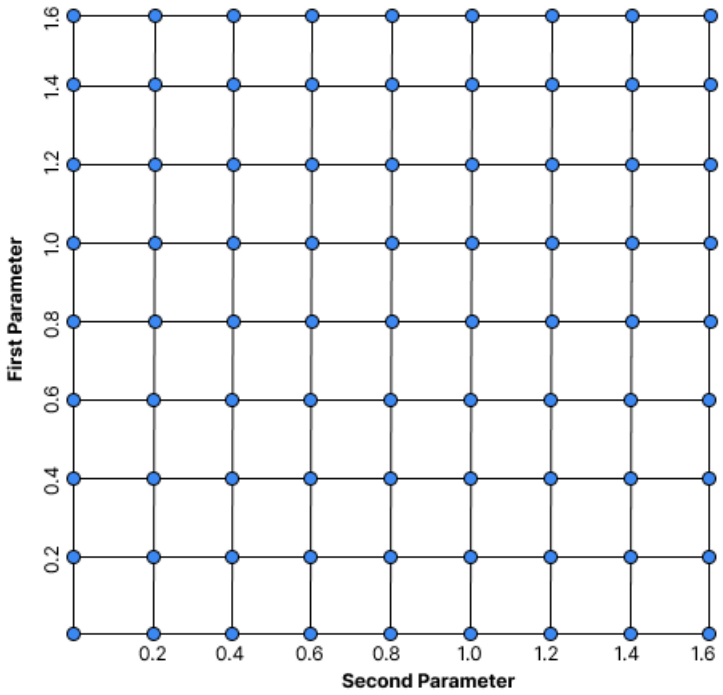

Figure 2.14 Illustration of grid search, where a grid with a range of possible parameters is set and searched over to achieve the best performance. Adapted from [129]

2.4 Databases, tools, and methods for cancer disease

In this section, I describe the databases, tools, and pipelines present in the literature, which have been used for a variety of analyses on cancer and CRC in particular.

2.4.1 Databases

Improvements in molecular biology technologies have created a large volume of biological data through several projects to analyze DNA, RNA, and proteins of many organisms around the world [130, 131]. Databases containing the output data of these studies can be found on the web, some with open access. According to Xiang et al. [132], biological databases can be divided into three categories: primary databases, which contain archives of raw sequence or structural data; secondary databases, which contain computationally processed or manually curated information, based on original information from the primary databases; and specialized databases, which contain information of particular research interest.

ncRNA databases

The ncRNA databases aim to store and organize data with relevant information on ncRNAs, which can be used by researchers to identify new sequences and in analysis [80]. Some ncRNAs databases are:

- NONCODE [133] has data collected from three sources: literature mining, GenBank, and specialized databases, such as: lncRNAdb [134] and LNCipedia [135] (`http://www.noncode.org/browse.php`);
- RFAM [79] contains ncRNA sequences and secondary structures. The RFAM data varies from data obtained through the use of multiple alignments, annotations of secondary structures, and covariance models (`https://rfam.xfam.org/`);
- miRbase [136] contains sequences, annotation and predictions of microRNAs (`http://www.mirbase.org/`);

Cancer Databases

As said before, ncRNAs play several roles in organisms, and interactions among these ncRNAs, such as lncRNAs and miRNAs, can influence disease regulation. Some specialized databases that focus on storing information on ncRNAs related to disease, and to cancer in particular, are:

- LncRNADisease [137] contains around 480 entries of experimentally supported associations between lncRNA and disease, including ncRNA interacting partners such as RNAs, miRNAs, and DNA (`http://www.cuilab.cn/lncrnadisease`);
- Lnc2Cancer [68] is a manually curated database that provides associations between lncRNA and human cancer (`http://www.bio-bigdata.net/lnc2cancer/`), experimentally supported;

- miRCancer [62] contains a collection of miRNAs related to human cancer, which are automatically extracted from articles in PubMed (`http://mircancer.ecu.edu/`);
- Gene Expression Omnibus DataSets portal (GEO) [138] stores original submitter-supplied records (series, samples, and tools) from NCBI articles, as well as curated datasets. The available search tool allows for filtering by cancer related files (`https://www.ncbi.nlm.nih.gov/gds`);
- The Cancer Genome Atlas (TCGA) [139] is a project that aims to catalog and discover major cancer-causing genomic alterations. Although it is not a database focused on ncRNAs, the TCGA contains several types of cancer related data, including sequencing reads, lincRNAs, and miRNAs (`https://portal.gdc.cancer.gov/`).

2.4.2 Tools

Given the biological aspects of the ncRNAs, there are several computational tools that use these features to predict, analyze, annotate and infer functions to these genes. These tools can be used alone or in conjunction with other tools, depending on the research goal.

ncRNA interactions and disease emergence

Most of the research that correlates ncRNAs interactions (mainly miRNAs and lncRNAs) with disease emergence is biological and focuses on the "decoy" effect [140, 30]. Most of the computational tools used in these biological studies are those that point to the miRNA binding site, which is used to identify whether a specific lncRNA can work as a decoy for the specific miRNA. These bioinformatic tools are:

- TargetScanS [141] (`http://www.targetscan.org/vert_72/`), which predicts biological targets of miRNAs by searching for conserved 8mer, 7mer, and 6mer sites that match the seed region of each miRNA;
- miRanda-mirSVR [142] (`http://www.microrna.org/microrna/getDownloads.do`), which uses a ML method to rank microRNA target sites by a down-regulation score;
- StarBase [143] (`http://starbase.sysu.edu.cn/starbase2/index.php`), which s decodes the interaction networks of lncRNAs, miRNAs, competing endogenous RNAs(ceRNAs), RNA-binding proteins (RBPs) and mRNAs from large-scale CLIP-Seq data.

2.4.3 Biological and computational methods related to CRC

In order to discuss the literature review of biological and computational methods related to CRC, I used the following three points as analysis criteria: (i) the research output of the study is a database with ncRNAs related to CRC; (ii) the study presents features associated to ncRNAs related to CRC, where these features can be used to build a ML model; and (iii) the study presents ncRNAs or experimentally proven interactions among proteins and ncRNAs, related to CRC.

Biological methods

The following studies present only biological methods and some used statistical analysis programs to generate information. Next, I describe these papers.

Han et al. [144] investigated the significance and biological function of the lncRNA UCA1 in CRC. Authors used biological methods and tools such as: RT-qPCR kits to evaluate the expression of UCA1 from tissue samples or cells; siRNAs designed from siRNA Construct software, in order to silence UCA1; cell proliferation assay, which evaluates the proliferation; and cell invasion assay[6]. Results correlated a high level of UCA1 expression to larger tumor size and suggested that UCA1 may regulate the expression of multiple genes influencing cancer cell proliferation, cell cycle progression, and apoptosis. Regarding the analysis criteria: (i) no output database was found; (ii) no ncRNA feature could be used; and (iii) the lncRNA UCA1 was experimentally confirmed as associated with CRC.

Zhong et al. [145] investigated the sponge effect of the lncRNA NEAT1 and miR-196a-5p. First, authors performed a study to relate mir-196 and CRC, where the results showed that miR-196a-5p inhibited both the protein level and mRNA level of the GDNF protein, which is related to cancer. After showing the role of mir-196 in CRC, the authors showed that NEAT1 acts as a decoy of mir-196 by using RT-qPCR, cell migration assay, cell viability assay[7] and through statistical analysis. In summary, the study found that NEAT1 inhibited the cell proliferation and migration potential of colorectal cells by acting as a miR-196a-5p decoy. Regarding the analysis criteria: (i) no output database was found; (ii) no ncRNA feature could be used; and (iii) the lncRNA NEAT1 and the miRNA miR-196a-5p were experimentally confirmed as associated with CRC.

Zhou et al. [74] investigated the relationship between the lncRNA HAND2-AS1 and miR-1275 interaction with the regulation of the KLF14 protein. Understanding of this relationship revealed HAND2-AS1 as a novel suppressor of CRC by sponging miR-1275.

[6] A cell invasion assay enables quantification of *in vitro* cell migration towards a membrane or a layer of cells such as endothelial cells.

[7] The cell viability assay analyzes the ability of cells to maintain or recover themselves.

For the study, the authors selected CRC samples from 74 patients undergoing surgery at Wenzhou Center Hospital. To evaluate the role played by HNAD2-AS1, the authors performed a biological pipeline using: cell counting assays; cell invasion assay; RT-qPCR; xenograft assay[8]; luciferase reporter assay; and statistical analysis using Student's t-test and one-way ANOVA. Regarding the analysis criteria: (i) no output database was found; (ii) no ncRNA feature could be used; and (iii) the lncRNA HAND2-AS1 and the miRNA miR-1275 were experimentally confirmed as associated with CRC.

Lu et al. [146] looked into the relationship between the lncRNA XIAP-AS1 and CRC. For the study, the authors collected 75 cancer tissue samples, and corresponding adjacent normal tissues were obtained from patients who underwent surgery for colon cancer without preoperative chemotherapy or radiotherapy. The authors performed a biological pipeline using: RT-qPCR; western blot[9]; cell viability assay; cell invasion assay; and statiscal analysis using SPSS. Results indicated that XIAP-AS1 promoted cell growth and invasion by facilitating the Wnt/β-catenin pathway and EMT. Regarding the analysis criteria: (i) no output database was found; (ii) no ncRNA feature could be used; and (iii) the lncRNA XIAP-AS1 was experimentally associated with colon cancer.

Gao et al. [3] studied the role played by lncRNA CACS15 in resistance to oxaliplatin (OXA) (a medicine used for CRC treatment) by sponging miR-145 in CRC. In order to perform the study, the authors collected 48 cancer tissue samples and corresponding adjacent normal tissues from CRC patients who underwent surgery at Shanghai Tongji Hospital of Tongji University School of Medicine. As the goal of the authors was to find the relationship with OXA resistance, the patients were divided into two groups: 25 were classified as OXA-resistant and 23 were classified as OXA-sensitive, where OXA regimen was defined as the appearance of new lesions or tumor growth $> 30\%$ after 2 months of chemotherapy while tumor growth $< 20\%$ was defined as non-resistance to OXA. The authors performed a biological pipeline using: cell transfection; RT-qPCR; OXA sensitivity assay; flow cytometric analysis; luciferase reporter assay; RNA pull-down assay[10]; RNA immunoprecipitation (RIP) assay[11]; western blot; and lentivirus production and infection. In the end, the study demonstrated that CASC15 knockdown enhanced OXA sensitivity in CRC cells by sponging miR-145 in CRC. Regarding the analysis criteria: (i) no output database was found; (ii) no ncRNA feature could be used; and (iii) the lncRNA CACS15 and the miRNA miR-145 were experimentally confirmed as associated with CRC and

[8]The xenograft assay serves to determine tumorigenic activity.

[9]The western blot is a technique to identify specific proteins from a complex mixture of proteins extracted from cells.

[10]RNA pull-down assay is a technique that enables the identification of proteins that interact with an RNA.

[11]The RNA immunoprecipitation (RIP) method shows the physical association between individual proteins and RNA molecules *in vivo*.

OXA resistance.

Ke et al. [147] investigated the role of miR-92a in the metastasis of CRC. To perform the study, authors used CRC tumor tissues from 158 patients who underwent surgical resection for CRC at the Department of Surgery, China Medical University Hospital. The authors performed a biological pipeline using: RT-qPCR; MiR-92a precursor and inhibitor transfection; cell invasion and migration assay; and statistical analysis using SPSS. Results demonstrated that miR-92a expression levels in the tumor tissues of CRC patients were positively correlated with nodal metastasis, where miR-92a promoted metastasis by suppressing PTEN gene expression and activating the PI3K/AKT pathway. Regarding the analysis criteria: (i) no output database was found; (ii) no ncRNA feature could be used; and (iii) the miRNA miR-92a was experimentally confirmed as associated with CRC.

Igarashi et al. [148] looked at the role of miR-31-5p in anti-EGFR therapy in CRC. Epidermal growth factor receptor (EGFR) is the target of anti-EGFR therapy, commonly used for treating CRC patients. The authors used 102 primary tumors of CRCs of patients who underwent surgical treatment and chemotherapy with anti-EGFR antibodies at Sapporo Medical University Hospital and Keiyukai Sapporo Hospital. All patients underwent surgical resection of primary CRC tumors before receiving anti-EGFR therapy. The authors performed a biological pipeline using: RT-qPCR and statistical analysis with JMP software, and found high miR-31-5p expression to be associated with survival in patients with metastatic CRC who underwent surgical treatment and chemotherapy with anti-EGFR. Regarding the analysis criteria: (i) no output database was found; (ii) no ncRNA feature could be used; and (iii) the miRNA miR-31-5p was experimentally associated with CRC.

Inoue et al [48] investigated the clinical significance and biological function of miR-29b in CRC. To perform the study, authors used tissues from 245 patients who underwent primary tumors at Osaka University Hospital, and performed: RT-qPCR, to quantify miR-29b expression; proliferation assay; flow cytometry; western blot; and statistical analysis with JMP10 software. Results indicated that miR-29b may be a novel prognostic marker and may play important roles in regulating tumor progression in CRC. Regarding the analysis criteria: (i) no output database was found; (ii) no ncRNA feature could be used; and (iii) the miRNA miR-29b was experimentally confirmed as associated with CRC.

Wang et al. [47] explored the role of serum miR-135a-5p and the potential of this RNA as a biomarker for diagnosis of CRC. To perform the study, authors used samples from 60 patients with primary CRC, 40 patients with CRC polyps, and 50 healthy controls, and performed: serum RNA extraction; cDNA synthesis; and statistical analysis using

SPSS. Through the biological pipeline, the authors detected that miR-135a-5p expression was elevated in the serum of CRC patients, and identified it as a potential biomarker for the diagnosis of CRC. Regarding analysis criteria: (i) no output database was found; (ii) no ncRNA feature could be used; and (iii) the miRNA miR-135a-5p was identified as a potential biomarker for CRC diagnosis.

Zu et al. [149] investigated the effects of miR-506 in CRC. The authors used data from the CRC cell lines SW480, SW620, HCT-116, and HT-29, which were cultured in DMEM medium (Gibco, USA)[12], and applied biological methods and tools such as: RT-qPCR; cell viability assay; colony formation assay; cell invasion and migration assay; and statistical analysis using SPSS. Authors found that miR-506 acts as a tumor suppressor in CRC by directly targeting LAMC1. Regarding the analysis criteria: (i) no output database was found; (ii) no ncRNA feature could be used; and (iii) the miRNA miR-506 was experimentally confirmed as associated with CRC.

Ozawa et al [150] evaluated the clinical significance of lncRNAs mapped in 8q24.21, the genomic region known as the gene desert in CRC. The 8q24.21 region is known for a lack of PC genes, which suggests the potential impact of lncRNAs in CRC. To evaluate significance of the lncRNAs, authors used 280 CRC and 20 adjacent tissues of patients from three institutes (Mie University, National Cancer Center Hospital, and Tokyo Medical and Dental University), and applied RT-qPCR assays to analyze the expression of 12 lncRNAs (PCAT1, PRNCR1, PCAT2, CCAT1, CCAT1-L, CASC19, CCAT2, CASC21, CASC8, CASC11, PVT1, and CCDC26). Using the results of differential expression and ROC curves with Youden's Index, the authors established optimal cut-off values for each lncRNA, as related to relapse-free survival (RFS)[13] and overall survival (OS)[14]. Findings showed that expression of CCAT1, CCAT1-L, CCAT2, PVT1, and CASC19 were elevated in cancer tissues, and high expression of CCAT1 and CCAT2 was significantly associated with poor RFS and OS, indicating them as potentially useful biomarkers for predicting tumor recurrence or CRC prognosis. Regarding the analysis criteria: (i) no output database was found; (ii) no ncRNA feature could be used; and (iii) the expression of lncRNAs CCAT1, CCAT1-L, CCAT2, PVT1, and CASC19 was shown to be elevated in CRC.

Dou et al. [151] investigated whether at decrease in lncRNA HOTAIR expression would inhibit CRC stem cells. Authors extracted data for this study from the human CRC LoVo cell line[15], and used biological methods and tools such as: RT-qPCR; proliferative assay;

[12]Dulbecco's Modified Eagle Medium (DMEM) is used for supporting the growth of many different mammalian cells.

[13]Measuring the relapse-free survival is one way to see how well a new treatment works.

[14]The percentage of people in a study or treatment group who are still alive for a certain period of time after they were diagnosed with or started treatment for a disease.

[15]Derived from a colon metastatic tumor, it is used to assess cancer immunotherapy agents *in vitro*.

colony forming assay; cell migration assay; cell invasion assay; western blot; and statistical analysis through a two-tailed paired Student's t-test. Results demonstrated that down-regulation of the HOTAIR expression in CRC decreases potential of tumorigenesis and metastasis. Regarding the analysis criteria: (i) no output database was found; (ii) no ncRNA feature could be used; and (iii) the lncRNA HOTAIR was found to be related to CRC.

Dou et al. [45] evaluated the role of miR-223 in resistance to doxorubicin, a drug used in CRC treatment. To perform the study, authors collected data from 50 paired CRC tissues and adjacent normal tissues from patients from Xianning Central Hospital, and used biological methods and tools such as: cytotoxicity assay; western blot; RT-qPCR; miRNA transfection; EGFP reporter assay; and statistical analysis through GraphPad Prism software. Results suggested that miR-223 promotes the doxorubicin resistance of CRC cells by targeting the FBXW7 protein. Regarding analysis criteria: (i) no output database was found; (ii) no ncRNA feature could be used; and (iii) the miRNA miR-223 and the protein FBXW7 were experimentally confirmed as associated with CRC.

Ma et al. [152] investigated the roles of the lncRNA BANCR and CSE1L gene in CRC. The author used 32 pairs of CRC tumor tissues and adjacent normal tissues and applied biological methods and tools such as: cell transfection; RT-qPCR assay; western blot; luciferase assay; MTT assay[16]; cell apoptosis assay; matrigel invasion assay; and statistical analysis using one-way ANOVA and Student's t-test. Results showed that BANCR silencing makes the CRC progression harder and enhances ADR, which is a drug used for CRC treatment, by regulating miR-203/CSE1L. The authors stated that targeting BANCR may be potentially therapeutic for CRC management. Regarding the analysis criteria: (i) no output database was found; (ii) no ncRNA feature could be used; and (iii) the miRNA miR-203 and the lncRNA BANCR were experimentally confirmed as associated with CRC.

Lee et al. [153] aimed to identify differentially expressed lncRNAs in 5-fluorouracil-resistant (a cancer drug) colon cancer cells. Authors extracted data from the cancer cell lines SNU-C4 and SNU-C5 and used biological methods and tools such as: MTT assay; RT-qPCR; flow cytometric analysis; and lncRNA profiling. The results suggested that the lncRNA snaR has a potential role as a negative regulator of cell growth in response to 5-fluorouracil-resistance. Regarding the analysis criteria: (i) no output database was found; (ii) no ncRNA feature could be used; and (iii) the lncRNA snaR was indicated to be involved in colon cancer drug resistance.

Yin et al. [154] explored the role of lncRNA GAS5 in CRC. Authors extracted data from 66 patients from the First Affiliated Hospital of Nanjing Medical University and

[16]The MTT is an assay for assessing cell metabolic activity.

used biological methods and tools such as: cell transfection; RT-qPCR; cell proliferation assays; tumor formation assays; and statistical analysis with SPSS. The results showed that overexpressed GAS5 could inhibit cell proliferation and demonstrated that GAS5 is downregulated in human CRC tumor tissues. Regarding the analysis criteria: (i) no output database was found; (ii) no ncRNA feature could be used; and (iii) the lncRNA GAS5 was experimentally confirmed to be associated with CRC.

Han et al. [155] examined the role of the lncRNA AFAP1-AS1 in tumor growth and metastasis of CRC. Authors extracted data from 15 patients from the Affiliates Hospital of Beihua University and used biological methods and tools such as: RT-qPCR; MTT assay; wound scratch assay; western blot; and statistical analysis by using SPSS. The results suggested that AFAP1-AS1 contributes to CRC tumor growth and metastasis. Regarding the analysis criteria: (i) no output database was found; (ii) no ncRNA feature could be used; and (iii) the lncRNA AFAP1-AS1 was experimentally associated with CRC.

Table 2.3 shows a summary of the biological methods developed for CRC found in the literature. Note that some of the studies do not present proteins, but rather interactions among lncRNAs, miRNAs and/or proteins. This is because the authors study the interaction with a protein that is confirmed to be related to CRC by previous studies.

Computational methods

Next, I describe another set of studies analyzed based on the use of biological data and tools, or biological databases along with a computational method.

Yuan et al. [156] studied the relationship among lncRNAs, miRNAs, and mRNAs in CRC. Authors extracted data from TCGA using a total of 480 CRC tumor tissues and 41 non-tumor tissues as input for bioinformatics prediction and correlation analysis; and authors selected 136 lncRNAs, 29 miRNAs, and 138 mRNAs to construct a lncRNA-miRNA-mRNA ceRNA network. The collected data also included clinical information such as age, gender, race, pathologic stage, pathologic tumor (pathologic T), pathologic node (pathologic N), and pathologic metastasis (pathologic M), noting that the authors filtered data based on expression level. The R package edgeR [157] was used to analyze the differential expression of mRNAs, miRNAs, and lncRNAs between the CRC tumor tissues and the normal samples. To identify the miRNA-lncRNA interactions, authors used data on experimentally verified miRNA-target genes from miRcode [158] and the miRTarBase, targetScan, and miRDB tools. To validate the miRNAs related to CRC, they used the miRCancer [62]. The final step of the study consisted in a Gene Ontology (GO) analysis using DAVID [159], in which authors also analyzed the correlation between the lncRNA-miRNA-mRNA and survival time by using the "survival" R package [160].

Table 2.3 Summary of biological methods, with observations regarding proteins, lncRNAs, miRNAs, and their interactions.

Paper	Biological methods	Proteins	LncRNAs	MiRNAs	Interactions
Han et al. [144]	RT-qPCR, cell proliferation assay and cell invasion assay	no	UCA1	no	yes
Zhong et al. [145]	RT-qPCR, cell migration assay and cell viability assay	no	NEAT1	miR-196a-5p	yes
Zhou et al. [74]	RT-qPCR, xenograft assay and luciferase reporter assay	no	HAND2-AS1	miR-1275	yes
Lu et al. [146]	RT-qPCR, western blot, cell viability assay and cell invasion assay	no	XIAP-AS1	no	yes
Gao et al. [3]	cell transfection, RT-qPCR,OXA sensitivity assay, flow cytometric analysis, luciferase assay, RNA pull-down assay, RIP and western blot	no	CACS15	miR-145	yes
Ke et al. [147]	RT-qPCR; cell invasion and migration assay	no	no	miR-92a	yes
Igarashi et al. [148]	RT-qPCR	no	no	miR-31-5p	yes
Inoue et al [48]	RT-qPCR, proliferation assay, flow cytometry and western blot	no	no	miR-29b	yes
Wang et al. [47]	serum RNA extraction and cDNA synthesis	no	no	miR-135a-5p	no
Zu et al. [149]	RT-qPCR, cell viability assay, colony formation assay, cell invasion and migrate assay	no	no	miR-506	no
Ozawa et al [150]	RT-qPCR	no	CCAT1, CCAT1-L CCAT2, PVT1 and CASC19	no	no
Dou et al. [151]	RT-qPC, proliferative assay, colony forming assay, cell migration assay, cell invasion assay and western blot	no	HOTAIR	no	no
Dou et al. [45]	cytotoxicity assay, western blot, RT-qPCR, miRNA transfection and EGFP reporter assay	FBXW7	no	miR-223	yes
Ma el al. [152]	cell transfection, RT-qPCR assay, western blot, luciferase assay, MTT assay, cell apoptosis assay and matrigel invasion assay	no	BANCR	miR-203	yes
Lee et al. [153]	MTT assay, RT-qPCR, flow cytometric and lncRNA profiling	no	snaR	no	yes
Yin et al. [154]	cell transfection, RT-qPCR, cell proliferation assays and tumor formation assay	no	GAS5	no	no
Han et al. [155]	RT-qPCR, MTT assay, wound scratch assay and western blot	no	AFAP1-AS1	no	yes

Results included the construction of, a ceRNA network based on lncRNA regulation, and the identification of lncRNAs LINC00400 and LINC00355 as promising therapeutic targets for CRC. Regarding the analysis criteria: (i) no output database was found; (ii) no ncRNA feature could be used; and (iii) LINC00400 and LINC00355 were found as novel lncRNAs associated with CRC.

Zhang et al. [8] aimed to investigate the clinical relevance and biological significance of the lncRNA IQCJ-SCHIP1 in CRC. Data came from 86 paired CRC tissues and adjacent tissues from patients of the Affiliated Hospital of Jiangnan University. In the biological steps, the authors used: RT-qPCR; transfection; CCK8 assay and colony formation assay; cell cycle and apoptosis analysis; and RNA-seq assay. After the biological steps, the authors used the clusterProfiler R package [161] to identify and visualize the Gene Ontology (GO) and enriched KEGG pathways of differentially expressed genes (DEGs). Then, they used the GSEA software to analyze DEGs [162]. Results identified IQCJ-SCHIP1 as a novel lncRNA that was down-regulated in CRC, indicative of a poor CRC prognosis. Regarding the analysis criteria: (i) no output database was found; (ii) no ncRNA feature could be used; and (iii) the lncRNA IQCJ-SCHIP1 was found as a potential therapeutic

target and prognostic factor for CRC.

Falzone et al. [163] aimed to identify differentially expressed miRNAs in CRC. The authors used datasets of microRNA profiling from the Gene Expression Omnibus DataSets portal (GEO DataSets) publicly available at NCBI. They filtered the GEO datasets by selecting only those containing: (i) information on both CRC patients and healthy patients as controls; and (ii) miRNA expression data of at least 30 samples. To perform the differential analysis of miRNAs, Authors used GEO2R tool, then, with the output from GEO2R, performed a statistical analysis to select the top 20 most significant up or down-regulated miRNAs in CRC. With the list of the top 20 for each dataset, the researchers used the bioinformatics tool Venn Diagrams, from the Bioinformatics and Evolutionary Genomics (BEG) to compare the sets. They divided miRNAs according to their expression levels: "highly", "moderately", "lightly" and "poorly" up-regulated or down-regulated then consulted the Catalogue of Somatic Mutation in Cancer (COSMIC) to identify the 10 most mutated genes that are known to be involved in CRC (APC, TP53, KRAS, FAT4, TGFBR2, LRP1B, PIK3CA, KMT2C, ZFHX3, BRAF). Next, authors used the bioinformatics tool microRNA Data Integration Portal (mirDIP) to evaluate the interaction between the miRNAs and the genes, then used DIANA-mirPath tool to identify the genes and pathways targeted by specific miRNAs and the statistical significance of this interaction. Results showed that hsa-miR-21-5p, miRNAs hsa-miR-195-5p and hsa-miR-497-5p target the highest number of genes within the pathways reported in CRC and may have an impact on CRC development. Regarding the analysis criteria: (i) A list of miRNAs and genes related to CRC can be found; (ii) no ncRNA feature could be used; and (iii) no ncRNAs were experimentally associated with CRC.

Bohme et al. [164] investigated the roles of miRNAs derived from cancer-associated fibroblasts (CAF), which is a cell type within the tumor that can modulate tumor progression in CRC. Authors extracted data from tumor patients who participated in an ongoing study developed at the UK National Institute of Health Research Clinical Research Network. The researchers used methods and tools such as: extraction of primary fibroblasts; isolation of exosomes; nanoString miRNA profiling; nanoString data analysis; RT-qPCR; miRNA pathway analysis, using KEGG and Ingenuity Pathway Analysis microRNA Target Filter (QIAGEN), and predicted mRNA targets using a combination of TargetScan, TarBase, miRecords, and the Ingenuity Knowledge Base; western blot; chemoresistance assay; proliferation assay; and statistical analysis with Student's t-test. Results exposed novel miRNA signatures specific to CAFs and presented miR-21 as an important molecule in CRC progression. Regarding the analysis criteria: (i) a list of miRNAs and genes related to CRC can be found; (ii) no ncRNA feature could be used; and (iii) the miRNA miR-21 was experimentally confirmed as associated with CRC.

Hu et al. [165] aimed to develop a lncRNA signature to improve CRC prediction. Authors downloaded data from the publicly available GEO databases. Because they were interested in the survival status of the patients, authors applied a filter to remove all samples without survival status, and they used samples from 895 patients. To analyze the expressions of lncRNAs, researchers used the GATExplorer tool and customized R scripts. In terms of risk of cancer, they used the GSEA tool to determine if patients of a given gene set were generally associated with risk score. To find whether or not lncRNAs within amplified (or deleted) regions affected expression levels, the statistical Mann-Whitney U (MWU) test was used. Next, the authors used cell viability assays and tumor cell invasion assays. Results showed that the lncRNAs AK123657, BX649059 and BX648207 were significantly down-regulated in CRC tissues, compared to normal colorectal tissues, suggesting a protective role in CRC biogenesis. Regarding the analysis criteria: (i) no output database was found; (ii) no ncRNA feature could be used; and (iii) the lncRNAs AK123657, BX649059 and BX648207 were associated with CRC.

Thorenor et al. [46] investigated the expression of disease-associated lncRNAs in CRC. The authors used 119 tissues from CRC patients who underwent surgery at Masaryk Memorial Cancer Institute and applied biological methods and tools such as: RT-qPCR; northern blot[17]; cell viability assay; cell cycle analysis; and western blot. After using biological tools, the study identified ZFAS1 as significantly up-regulated, and then used four bioinformatic algorithms: miRanda, miRWalk, RNA22, and Targetscan to predict miRNAs sponged by ZFAS1. Finally, authors performed a statistical analysis using R scripts with packages of Bioconductor [166] and LIMMA approach, combined with hierarchical clustering (HCL) to evaluate survival rate. Results demonstrated that ZFAS1 can act as a sponge of miR-150-5p, influencing the metastatic potential and miR-590-3p, which affects the CDK1 protein previously related to cell cycle and proliferation. Regarding the analysis criteria: (i) no output database was found; (ii) no ncRNA feature could be used; and (iii) the lncRNA ZFAS1 and the miRNAs miR-590-3p and miR-150-5p were associated with CRC.

Qiu et al. [167] investigated the correlation between the expression of lncRNAs and CRC. The study used data from published datasets in the gene expression omnibus (GEO). The authors filtered the datasets based on the following conditions: (i) patients with CRC; (ii) CRC tissue and normal tissue samples were available for comparison; (iii) data were obtained from the same platform; and (iv) there were more than three samples. In order to identify the differentially expressed probe sets, authors adopted a workflow based on Yang et al. [168], then they reannotated the datasets by downloading

[17]The northern blot is a technique used to evaluate gene expression by detection of RNA (or isolated mRNA) in a sample.

the sequences of the probes from the official Affymetrix website and executed BLAST in order to deepen their analysis. After reannotating and filtering the dataset, the authors used the obtained lncRNAs as input for a Principal component analysis (PCA) using the R Bioconductor package and performed a hierarchical clustering analysis (HCA) on the differentially expressed lncRNAs. Combining HCA and PCA, the authors selected 25 differentially expressed lncRNAs, which were chosen as markers. The results indicated that CRC tissue could be discriminated from tumor-adjacent normal tissue with an accuracy of 88.8% (71 out of 80 samples were correctly classified). Despite this accuracy of 88.8%, the authors did not present the criteria used to establish the 25 lncRNAs as features. Regarding the analysis criteria used to analyze this paper: (i) a list of lncRNAs that may be related to CRC can be found; (ii) no ncRNA feature could be used; and (iii) no ncRNAs were experimentally associated with CRC.

Gründner et al. [9] proposed the use of machine learning and of biological and clinical features to predict chemotherapy use in patient treatment. The authors used data from 564 colon and rectal cancer patients from Erlangen University Hospital. First, they separated the data into two groups: patients who received chemotherapy and patients who did not receive it. Second, they compiled a list of biological features (a list of 59 genes) and clinical features: cancer localization; patient gender; whether the patient is a smoker; patient weight; patient height; cancer type; and tumor stage. Third, with the selected features, authors built ML models with: DT; RF; and deep neural networks. Finally, their best model, which uses RF and combines all clinical and biological features showed an accuracy of 71% as outcome. Regarding the analysis criteria: (i) no output database was found; (ii) the list of biological features was not provided, but a list of clinical features was provided; and (iii) no ncRNAs were experimentally associated with CRC.

Gupta et al. [11] described an approach using ML and clinical features to predict colon cancer stages and the survival period of the patient. First, the authors extracted data from 4,021 patients from Chang Gung Memorial Hospital. Then, they selected the features: body mass; family history of cancer; age; gender; smoking and alcohol consumption; hemoglobin level; creatinine level; and white blood cells. With these features, the authors devised a pipeline using: SVM, LR, Multilayer Perceptrons (MLP), KNN, and AB to build the predictions model, which had an accuracy of approximately 84%. Regarding the analysis criteria: (i) no output database was found; (ii) there were no biological features were not provided, but a list of clinical features was provided; and (iii) no ncRNAs were experimentally associated with CRC.

Achilonu et al. [10] also described a pipeline to predict recurrence and patient survival using clinical features. First, the authors extracted data from 716 patients with CRC from Johannesburg Hospitals. Then, they selected the features: pathology; race; recur-

rence; chemotherapy; histology; and other clinical aspects. With the features, the authors devised a pipeline using: LR, Naive Bayes (NB), C5.0, RF, SVM, and Artificial Neural Network (ANN) to build the predictions model, which had an accuracy of approximately 87% at best. Regarding the analysis criteria: (i) no output database was found; (ii) there were no biological features were not provided, but a list of clinical features was provided; and (iii) no ncRNAs were experimentally associated with CRC.

Table 2.4 shows a summary of the computational methods found in the literature for the study of CRC. Note that some of the projects do not present a protein, but rather present interactions among lncRNAs, miRNAs, and/or proteins. This is because they study the interaction with a protein that was confirmed to be related to CRC in previous research.

Table 2.4 Computational methods summary

Paper	Computational methods	Proteins	LncRNAs	MiRNAs	Interactions
Yuan et al. [156]	miRTarBase, targetScan and R	no	LINC00400 and LINC00355	no	yes
Zhang et al. [8]	clusterProfiler	no	IQCJ-SCHIP1	no	no
Falzone et al. [163]	BEG	list of genes	no	list of miRNAs	yes
Bohme et al. [164]	DIANA-mirPath and QIAGEN	no	no	miR-21	no
Hu et al. [165]	GATExplorer and R	no	AK123657, BX649059 and BX648207	no	no
Thorenor et al. [46]	Bioconductor, miRanda, miRWalk, RNA22 and Targetscan	no	ZFAS1	miR-590-3p and miR-150-5p	yes
Qiu et al. [167]	Bioconductor, BLAST and HCA	no	list of lncRNAs	no	no
Gründner et al. [9]	C50, RF and DT	no	no	no	no
Gupta et al. [11]	SVM, LR, MLP, KNN and AB	no	no	no	no
Achilonu et al. [10]	LR, NB, C5.0, RF, SVM and ANN	no	no	no	no

Chapter 3

Competing endogenous RNAs in CRC

In this chapter, we present a method to predict ceRNAs and biological markers that can be used for CRC prognosis, which was published at Vieira et al [169]. In Section 3.1, we propose a pipeline to identify potential biological markers that can be used for CRC prognosis for the colon, rectum, and rectosigmoid junction. In Section 3.2, we present the results obtained for each CRC anatomical site, together with the predicted ceRNAs and biomarkers. In Section 3.3, we discuss the obtained results.

3.1 A method to predict biological markers

In this section, first, we describe the general pipeline developed to predict biological markers used for CRC prognosis and the data used as input. We also detail how the differential expression analysis is done and show how the ceRNA networks are constructed. Finally, we discuss the methods used to perform the functional and survival analysis of CRC.

3.1.1 General pipeline and input data

To identify biological markers used for CRC prognosis, we propose a pipeline (Figure 3.1) with four steps: differential expression (DE) analysis; ceRNA network construction; functional analysis; and survival analysis. This pipeline was defined for each of the three CRC anatomical sites: colon, rectum, and rectosigmoid junction.

As the input for our pipeline, and given that CRC itself is so heterogeneous, to analyze the patient markers prognosis with minimum variance from family, hereditary, or other outlier cases, we initially use data only from patients with adenocarcinoma. The

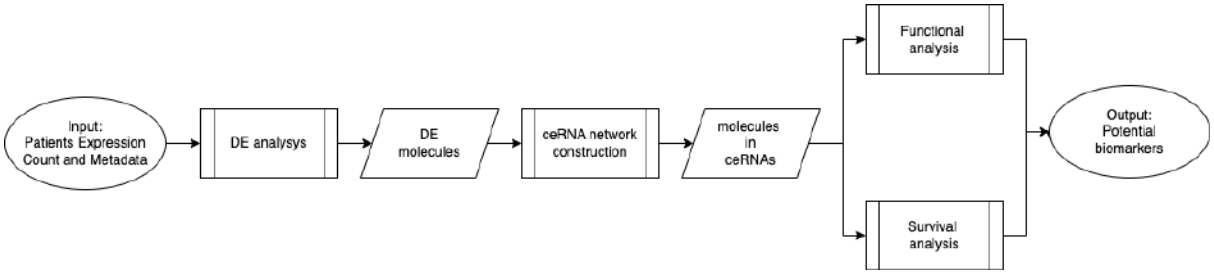

Figure 3.1 Pipeline to predict potential biomarkers related to CRC. Using both the RNA expression raw count data and the clinical metadata from patients as input, first we perform a DE analysis, then with the output we construct related ceRNA networks, and afterward, we perform an enrichment and functional analysis with the molecules present in the ceRNA networks.

needed biological and clinical information is extracted from The Cancer Genome Atlas (TCGA), a database created as a joint effort between NCI and the National Human Genome Research Institute, that characterized over 20,000 cancer samples of over 33 cancer types [59]. In specific, two projects from TCGA were used: TCGA rectal adenocarcinoma (TCGA-COAD); and TCGA rectal adenocarcinoma (TCGA-READ). From both projects, we collected RNA expression raw count data of 541 primary tumors and 48 non-tumor tissues from 539 patients, where 391 patients had cancer in the colon, 85 had cancer in the rectum and 69 had cancer in the rectosigmoid junction cancer. In the sequence, we explain each step of the pipeline (Figure 3.1).

3.1.2 DE analysis and ceRNA network construction

Using the RNA expression raw count data obtained from TCGA as input, we created a custom R script using the GDCRNATools v1.6 [170] package to perform a DE analysis for each CRC anatomical site. First, we normalized the data using Voom normalization and then used the limma [171] to compare primary tumor tissues against non-tumor tissues to obtain a list of DE miRNAs, lncRNAs, and Protein Coding (PCs). We just considered molecules with $FDR \leq 0.05$ and $|logFC| \geq 2$.

From the output of the DE molecules obtained, we also created a custom R script to generate the ceRNA networks for each CRC anatomical site and to analyze the differences and intersections among the networks. To generate the ceRNA networks, we also used the method from GDCRNATools that as general criteria infer the competing endogenous interactions between lncRNA and mRNA pairs if: the lncRNAs and mRNAs share a significant number of miRNAs; the lncRNA and mRNA expression are positively correlated; and the shared miRNAs play similar roles in regulating the lncRNA and mRNA expression. To check if the criteria are met, GDCRNATools use spongeScan [172] algorithm and

the starBase v2.0 [173] database. The ceRNA network construction script provides as output a list of nodes representing lncRNAs, miRNAs, and PC; and a list of interactions among the nodes.

3.1.3 Functional and survival analysis

With the output from the ceRNA network molecules and the patient clinical metadata merged, we created a R script to use this input to perform a survival and functional analysis. For the functional analysis, we performed an enrichment analysis from the input data against three pathways databases: Gene Ontology (GO) [174]; Kyoto Encyclopedia of Genes and Genomes (KEGG) [175] and Disease Ontology (DO) [176]. As a reference, the used annotation for humans was from the org.Hs.eg.db database v3.11.4. We also considered only the results with p-value ≤ 0.05. For further analysis, we also included some pathways presenting $FDR > 0.05$ as they can show good discussion points for the CRC functional analysis. For the survival analysis, we used two methods: Cox Proportional-Hazards (CoxPH) and the Kaplan Meier (KM). As an output of the analysis, the script gives a list of molecules that affect patient survival. As an output from CoxPH, we also extract the molecule hazard ratio (HR) information, in which HR > 1 indicates risk factors, and HR < 1 indicates protective factors. After calculating the HR we used the confidence intervals and removed the outliers by keeping only the molecules with $|higherLimit - HR| \leq 6$ and $|lowerLimit - HR| \leq 6$ were considered. As an output from KM we plotted the patient survival curves associated with each resulting molecule. For both algorithms, we used $p < 0.05$ as the threshold for the results.

3.2 Results

This section presents the results from the pipeline described in the previous sections. We show the results of the DE analysis for the colon, rectum, and rectosigmoid junction. Then, we present the resulting ceRNAs network for the colon, rectum, and rectosigmoid junction, highlighting their differences and intersections. After, we show the functional analysis results for the colon, rectum, and rectosigmoid junction. And next, we show the survival analysis results for all the anatomical sites.

3.2.1 DE molecules

As proposed in the pipeline, we performed a DE analysis for each CRC anatomical site. For the colon, we obtained a total of 3,649 DE molecules (Figure 3.2), where from these

we had: 1,179 up-regulated and 1,906 down-regulated PCs; 213 up-regulated and 136 down-regulated miRNAs; and 140 up-regulated and 75 down-regulated lncRNAs.

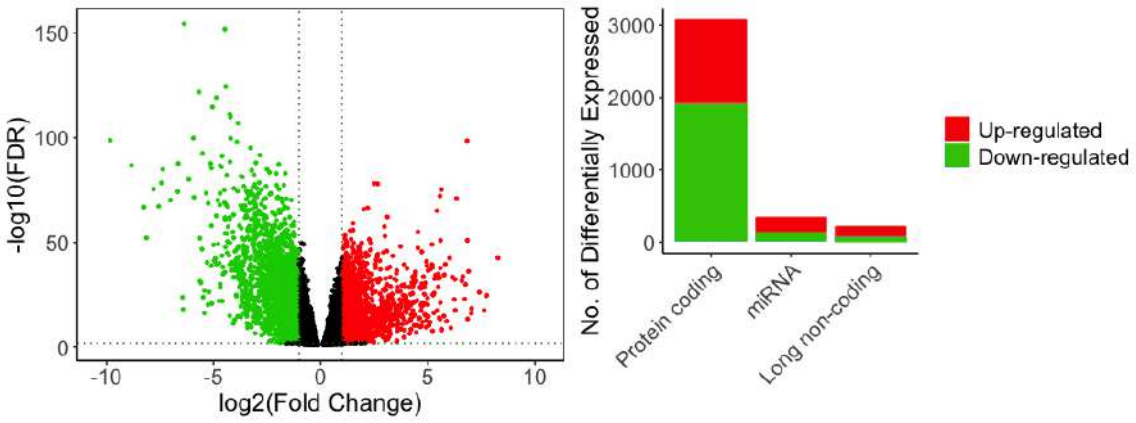

Figure 3.2 Volcano and bar plot with DE PCs, miRNAs, and lncRNAs from colon cancer patients. The total count of 3,649 DE molecules, where 1,532 were up-regulated and 2,117 were down-regulated.

For the rectum, we obtained a total of 2,368 DE molecules (Figure 3.3), where from these we had: 535 up-regulated and 1,532 down-regulated PCs; 119 up-regulated and 99 down-regulated miRNAs; and 46 up-regulated and 37 down-regulated lncRNAs.

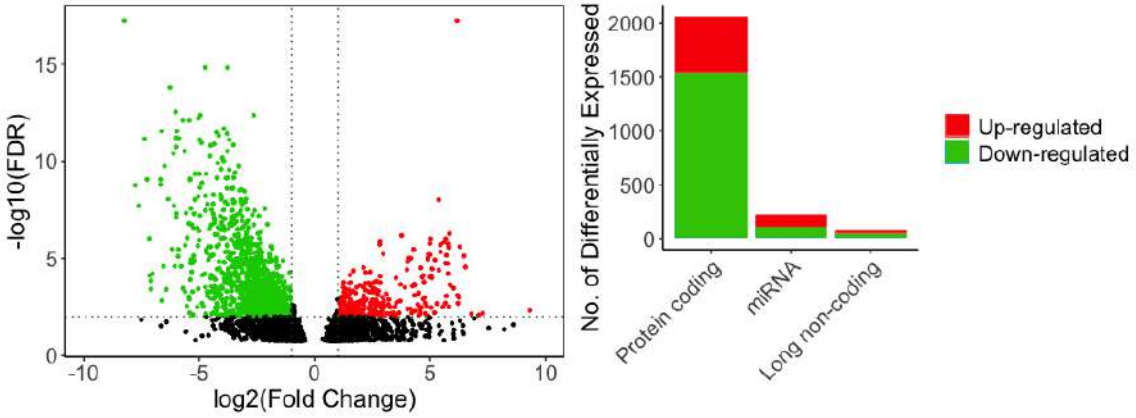

Figure 3.3 Volcano and bar plot with DE PCs, miRNAs, and lncRNAs from rectum cancer patients. The total count of 2,368 DE molecules, where 700 were up-regulated and 1,668 were down-regulated.

For the rectum, we obtained a total of 3,382 DE molecules (Figure 3.4), where from these we had: 1,005 up-regulated and 1,880 down-regulated PCs; 181 up-regulated and 108 down-regulated miRNAs; and 149 up-regulated and 59 down-regulated lncRNAs.

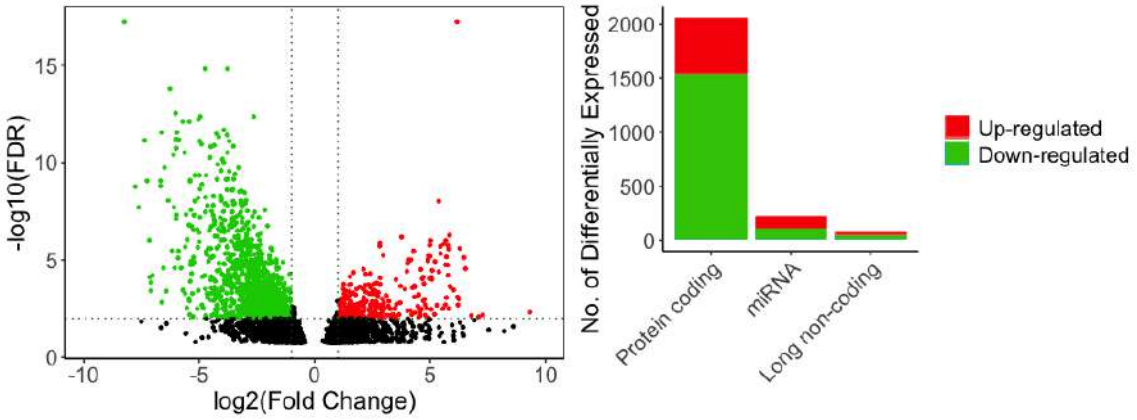

Figure 3.4 Volcano and bar plot with DE PCs, miRNAs, and lncRNAs from rectosigmoid junction cancer patients. The total count of 3,382 DE molecules, where 1,335 were up-regulated and 2,047 were down-regulated.

3.2.2 CeRNA Networks

After obtaining the DE molecules for the colon, rectum, and rectosigmoid junction, we used the DE molecules to generate the ceRNA networks. The ceRNA network is represented by a graph, where its nodes are molecules present in the network, and the lines connecting the nodes are the interactions of the molecules. In order to understand better the ceRNA behavior for each anatomical site, we built one network for each site and also explored their intersections.

For the colon, a ceRNA network consisting of 239 nodes and 506 interactions was established (Figure 3.5). From these 239 molecules, we had: 161 PCs, 60 miRNAs, and 18 lncRNAs. We can notice that in this network we also have a sub-network containing most of the interactions. This sub-network contains the lncRNA H19, which has been reported to have a protagonist role regulating various cancer-related PCs in colon cancer and in CRC in general [2, 7, 8].

For the rectum, a ceRNA network consisting of 82 nodes and 139 interactions was established (Figure 3.6). From these 82 molecules, we had: 70 PCs, 8 miRNAs, and 4 lncRNAs. In this network, we can also notice a sub-network containing most of the interactions. This sub-network contains the lncRNA MAGI2-AS3, which has been previously pointed out as related to cell apoptosis and proliferation in CRC [177].

For rectosigmoid junction, a ceRNA network consisting of 133 nodes and 212 interactions was established (Figure 3.7). From these 133 molecules, we had: 93 PCs, 26 miRNAs, and 14 lncRNAs. In this network, we can notice two subnetworks containing most of the interactions. One of these subnetworks also contains the lncRNA MAGI2-AS3, which is present in the biggest sub-network from the rectum and was pointed out with a regulation role in CRC development. The second of these subnetworks contains

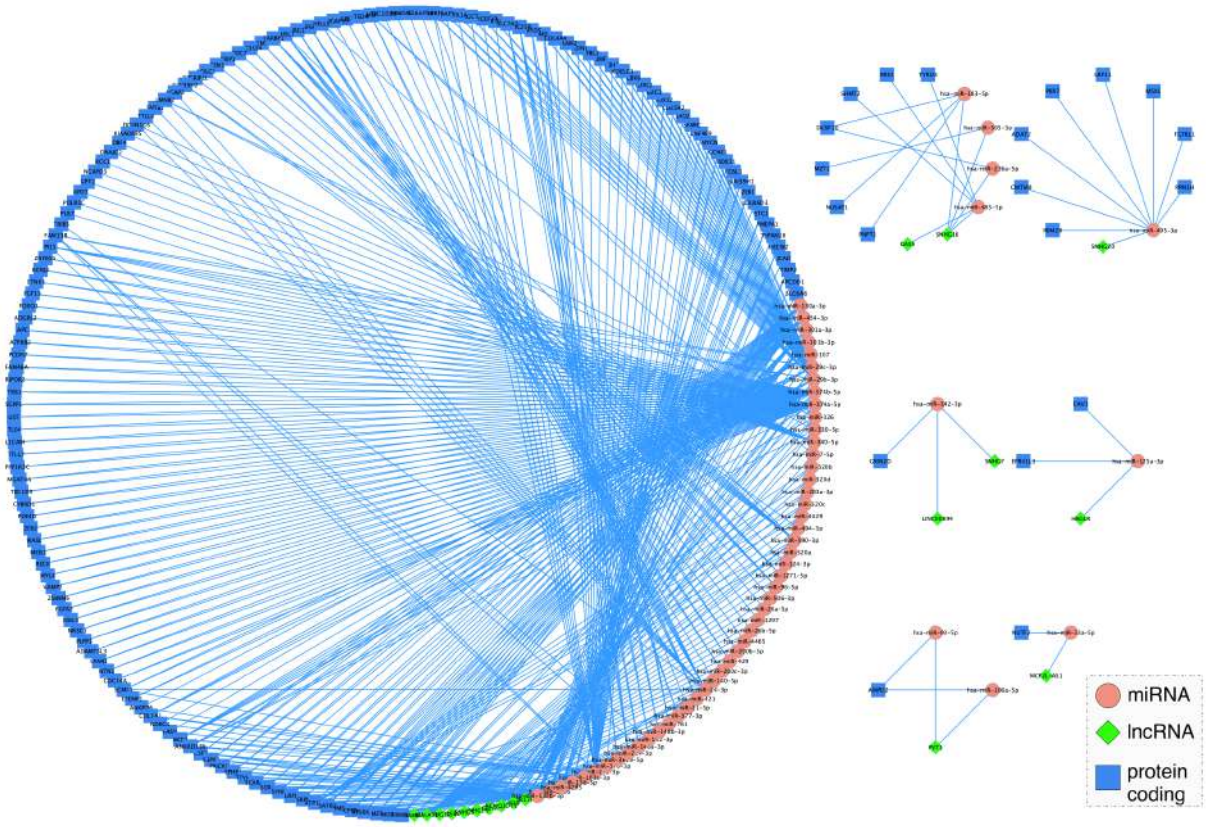

Figure 3.5 Colon ceRNA network with 239 nodes and 506 interactions. We can note 7 subnetworks present, where one of them contains most of the molecule's interactions.

the lncRNAs: SNHG1 and SNHG15, which were previously pointed as having an active role in cell proliferation, apoptosis, and activation of the Wnt/β-catenin signal in CRC [178, 179, 180, 181]

For the intersection between the colon and rectum ceRNA networks, we can notice a uniqueness of 2 nodes and 2 interactions, which are the PCs: HMMR and HELLS. HMMR was already pointed out a as potential regulation in CRC, but their specific roles need to be further clarified [182, 183]. For the intersection between the colon and rectosigmoid junction ceRNA networks, we can notice a uniqueness of 48 nodes and 77 interactions, which can indicate that the colon and rectosigmoid junction share more similarities than the colon and rectum. For the intersection between the rectum and rectosigmoid junction ceRNA networks, we can notice a uniqueness of 12 nodes and 23 interactions, also showing a possible indication that the rectum and rectosigmoid junction share more similarities than colon and rectum, thus possibly highlighting the problem of misdiagnosing colon and rectum cancer in the rectosigmoid junction, other than the anatomical proximity. For the intersection among all anatomical sites, we can notice a uniqueness of 47 nodes and 76 interactions, which can indicate the common mechanism in the regulation of

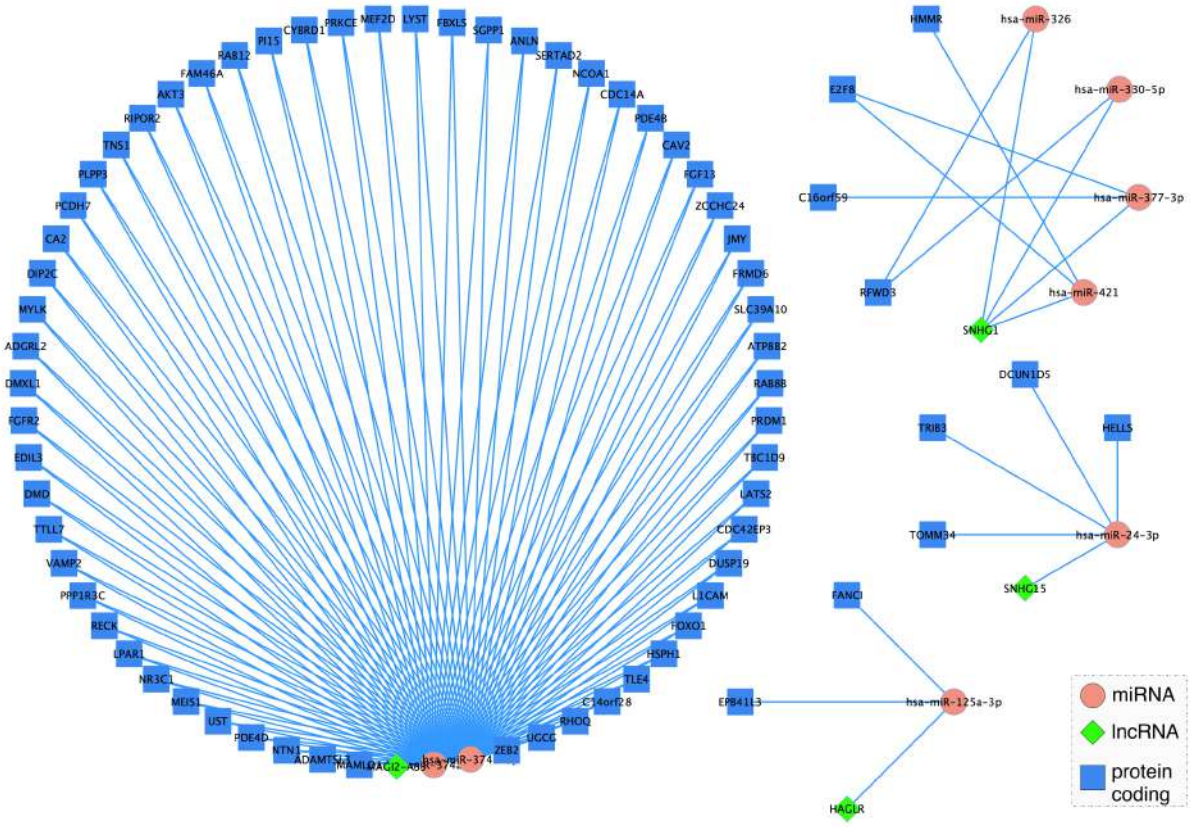

Figure 3.6 Rectum ceRNA network with 82 nodes and 139 interactions. We can note 4 subnetworks present, where one of them contains most of the molecule's interactions.

CRC at all anatomical sites. Finally, we also show the nodes and interactions unique to each anatomical site, where: the colon presents 142 nodes and 351 unique interactions; the rectum presents 18 nodes and 35 unique interactions; and the rectosigmoid junction presents 24 nodes and 34 unique interactions, showing the possible specific underlying mechanism to CRC development in each anatomical site. Figure 3.8 shows the described networks.

Figure 3.9 better illustrates the common mechanism of CRC in all anatomical sites, where we can notice four subnetworks, which of them contain most of the interactions. This sub-network contains the lncRNA MAGI2-AS3 and affects a great number of PCs by interacting with hsa-miR-374b-5p and hsa-miR-374a-5p.

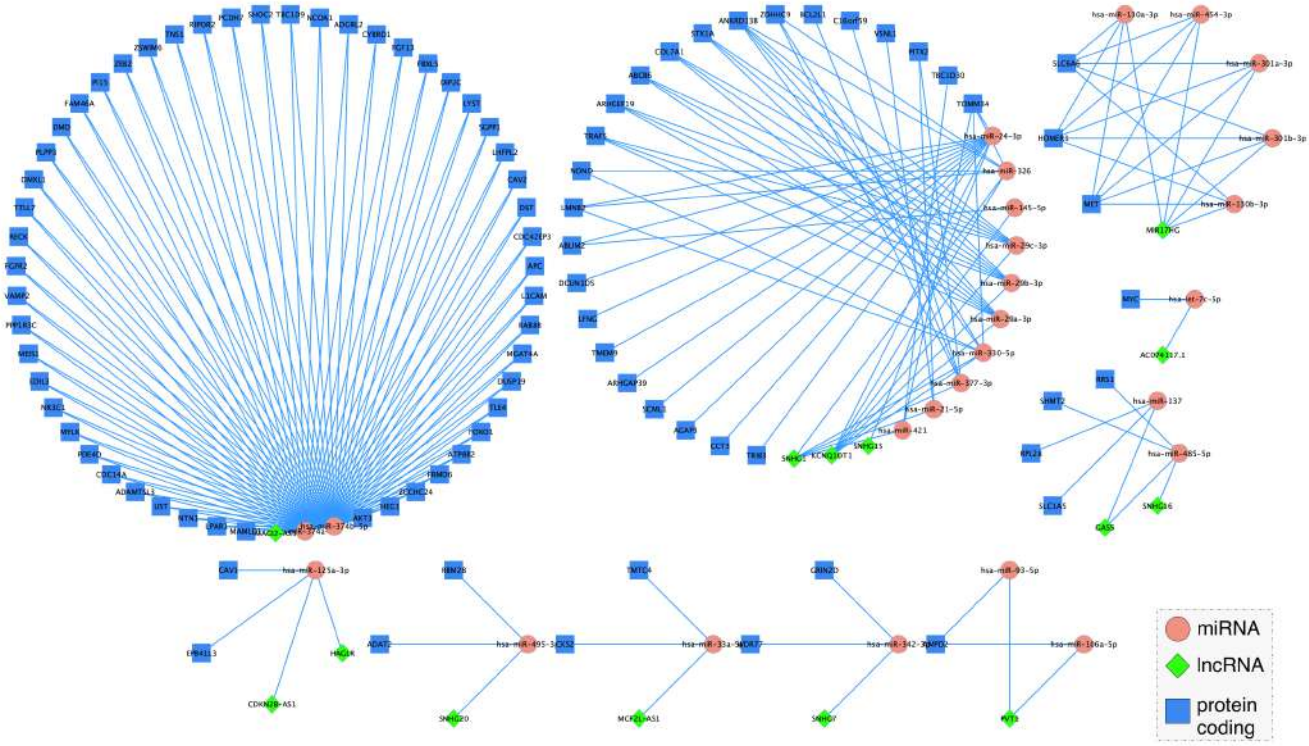

Figure 3.7 Rectosigmoid junction ceRNA network, with 133 nodes and 212 interactions. We can note 10 subnetworks present, where two of them contain most of the molecule's interactions.

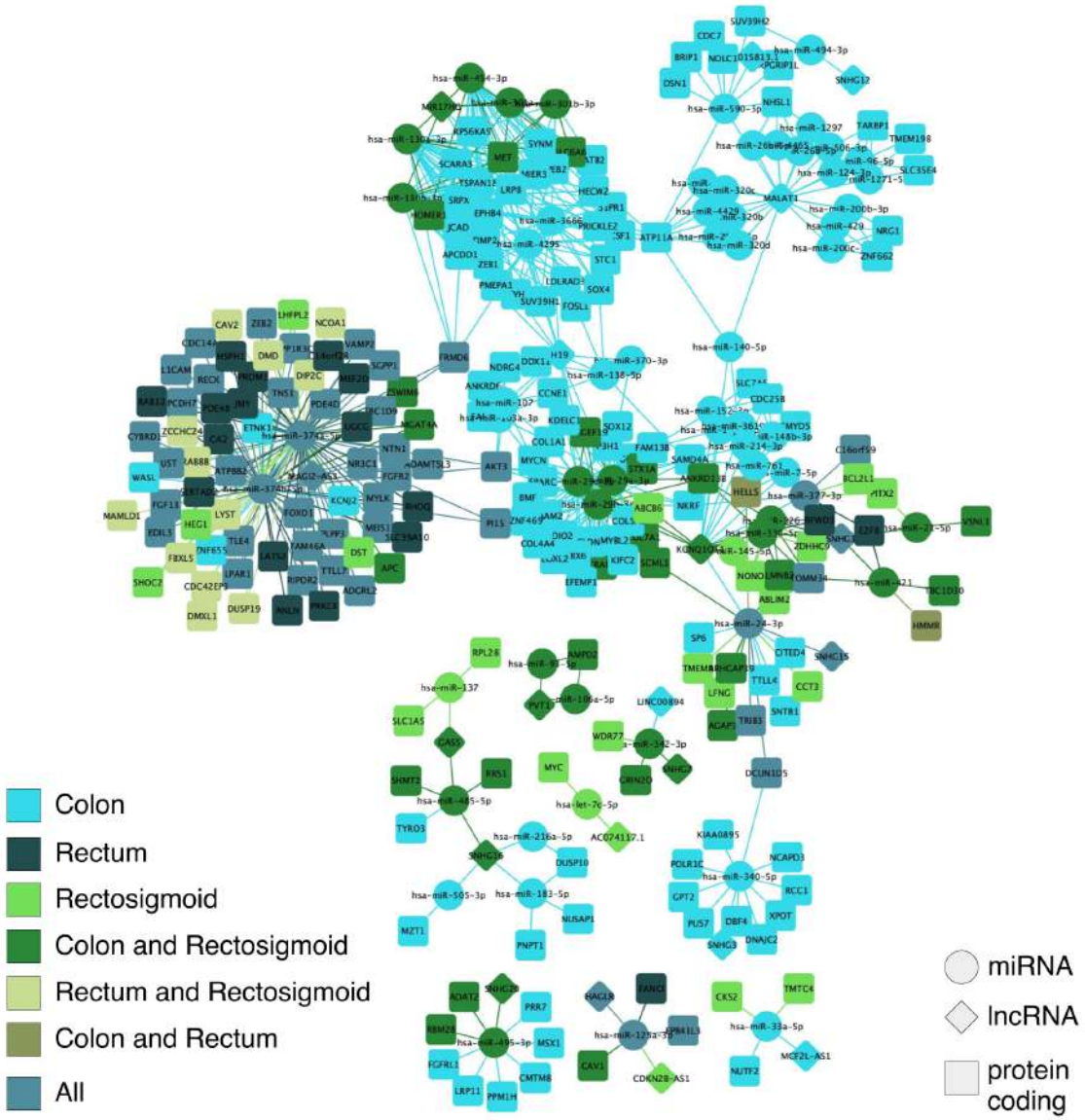

Figure 3.8 Competing endogenous RNA (ceRNA) network in colon, rectum, and rectosigmoid junction sites. The diamonds represent lncRNAs, the circles represent miRNAs, and the squares represent *PCs*. The molecules and interactions of each CRC site can be identified by color. Published at Vieira et al [169].

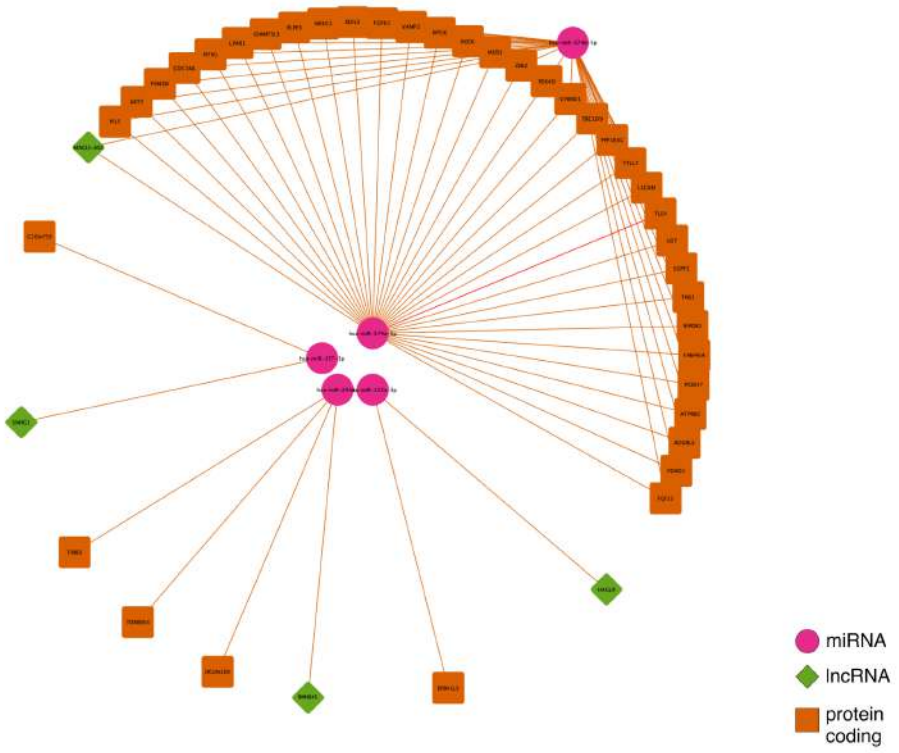

Figure 3.9 Competing endogenous RNA (ceRNA) network intersection for colon, rectum, and rectosigmoid junction sites. Published at Vieira et al [169].

3.2.3 Functional analysis

After obtaining molecules present in the ceRNA networks for the colon, rectum, and rectosigmoid junction, we used these molecules with the patient's clinical data to perform the functional analysis. For colon (Figure 3.10), we can see a heterogeneous enrichment for the different databases. The most interesting enrichment results come from KEGG, which relates the input molecule with miRNAs in cancer and prostate cancer pathway, and DO, which relates to central nervous system cancer pathways.

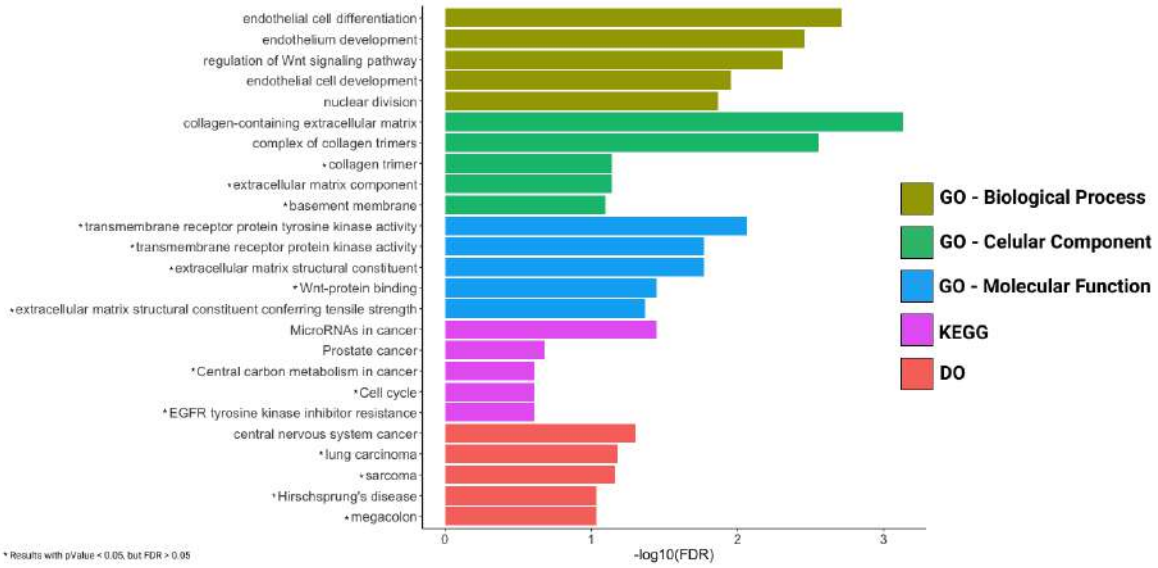

Figure 3.10 Functional enrichment analysis of molecules present in colon ceRNA network. The top 5 enrichment results for GO biological processes, cellular component, molecular function, DO and KEGG are shown in different colors. Asterisks (*) indicate pathways presenting $FDR > 0.05$.

For the rectum, we can see that the enrichment shows many pathways related to signaling (Figure 3.11). We can also notice the enrichment against KEEG shows a relationship with insulin resistance, which is interesting given the connection between CRC and the negative effect of diabetes on the patient overall survival.

For rectosigmoid junction, we can note that the enrichment shows many pathways related to signal transduction (Figure 3.12). From the enrichment against KEGG, we can also notice pathways related to the central carbon metabolism in cancer and small-cell lung cancer.

To understand the possible specific mechanisms that differentiate CRC progression in the different anatomical sites, we performed a functional analysis for the molecules that are unique in the colon, rectum, and rectosigmoid junction (Figure 3.13). For colon specific molecule enrichment, we can notice pathways related to nuclear division, endothelial cell differentiation, and collagen regulation. In fact, collagen has already been reported as an

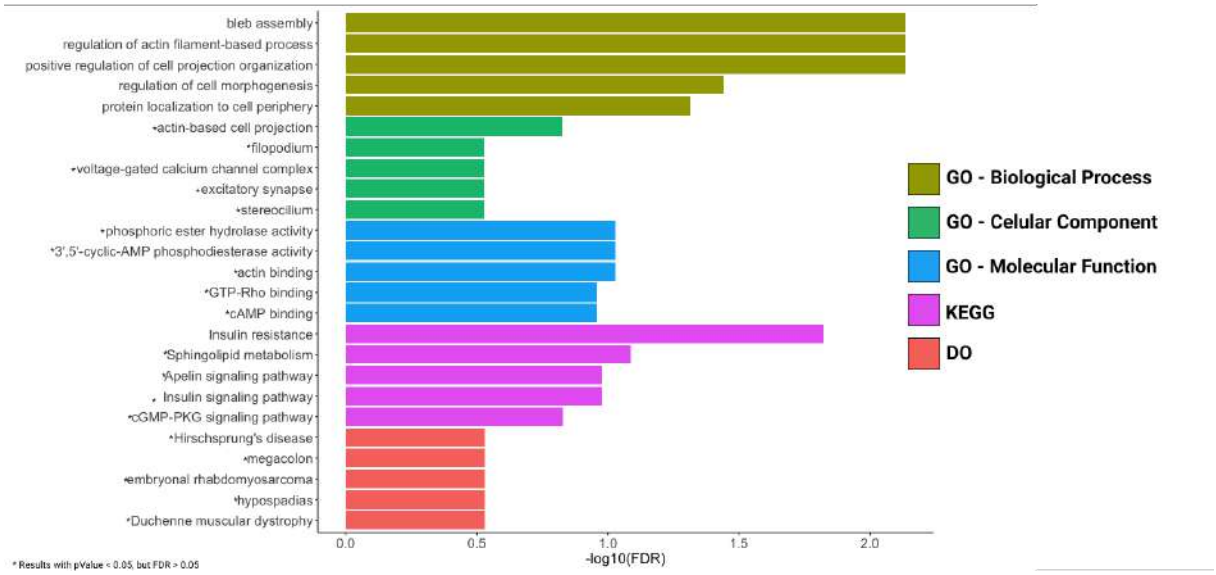

Figure 3.11 Functional enrichment analysis of molecules present in rectum ceRNA network. The top 5 enrichment results for GO biological processes, cellular component, molecular function, DO and KEGG are shown in different colors. Asterisks (*) indicate pathways presenting $FDR > 0.05$.

important factor in regulating cancer tumorgenesis in CRC [184, 185]. For rectum specific molecule enrichment, we can notice pathways related to cell differentiation and signaling. For rectosigmoid junction specific molecule enrichment, we notice a behavior similar to the one of the complete network.

Finally, we performed an enrichment analysis for the molecules present in the common ceRNA network of CRC anatomical sites (Figure 3.14). Again, we can notice the enrichment related to insulin resistance as noticed in the rectum enrichment. Also, we can notice known pathways related to cancer development such as cell proliferation and Wnt signaling.

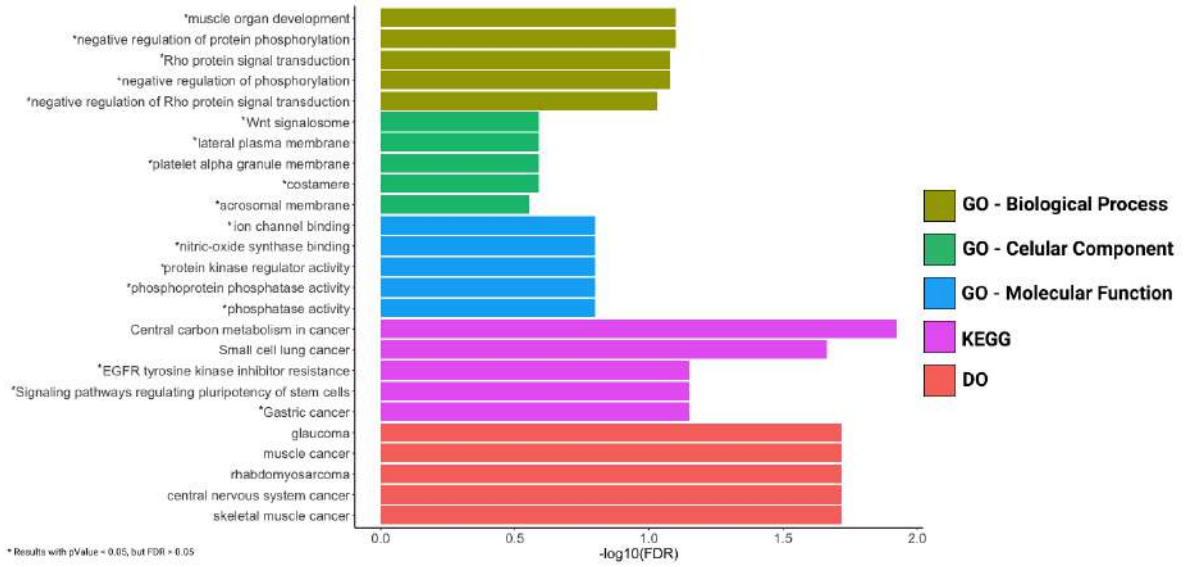

Figure 3.12 Functional enrichment analysis of molecules present in rectosigmoid junction ceRNA network. The top 5 enrichment results for GO biological processes, cellular component, molecular function, DO and KEGG are shown in different colors. Asterisks (*) indicate pathways presenting $FDR > 0.05$.

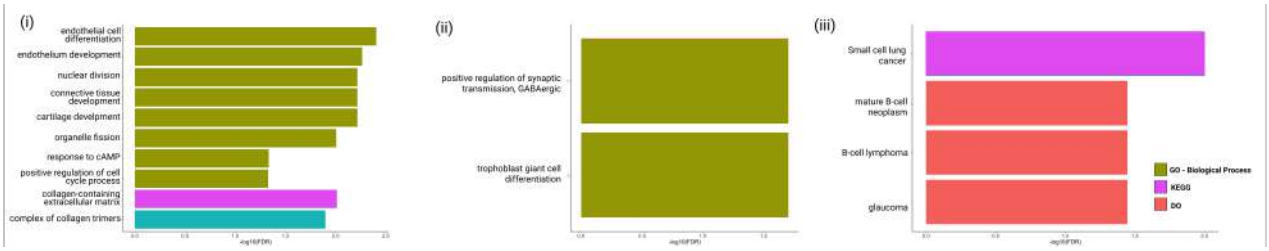

Figure 3.13 Functional enrichment analysis of molecules unique to the colon (i), rectum (ii), and rectosigmoid junction (iii) ceRNA networks.

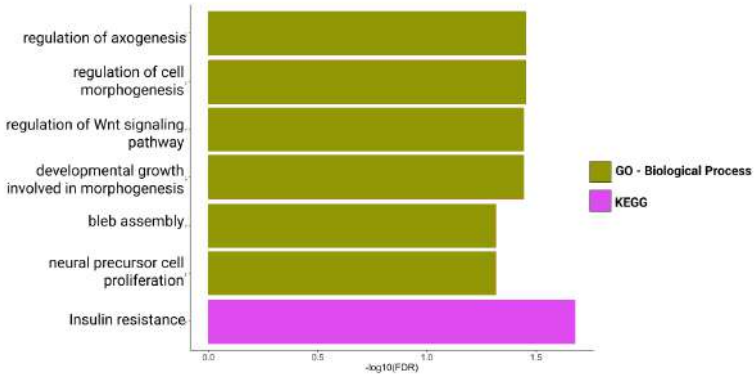

Figure 3.14 Functional enrichment analysis of molecules common to colon, rectum, and rectosigmoid junctionceRNA networks.

3.2.4 Survival analysis

Also using the molecules present in the ceRNA networks for the colon, rectum, and rectosigmoid junction and the patient's clinical metadata, we performed the survival analysis. By using the CoxPH method to calculate the HR for each CRC site (Figure 3.15), we identified 20 potential molecules that affect patient survival. Of these 20 molecules with relevant HR: 14 were from the colon; 3 from the rectum; 3 were from the rectosigmoid junction; and 1 was common in the rectum and rectosigmoid junction. DMD was the gene with the highest HR for rectum and rectosigmoid junction, while AGAP3 was the highest one for colon patients. It is important to notice that the colon displayed most of the molecules with relevant HR, which may happen because the data has more patients with CRC at the colon site.

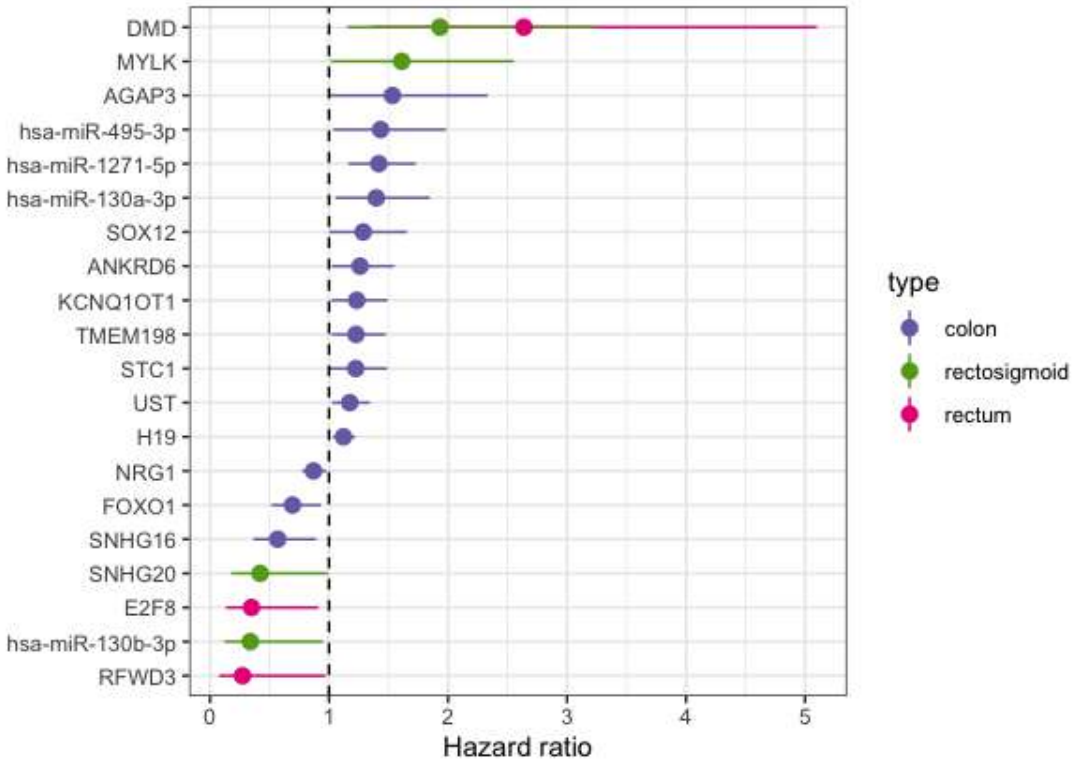

Figure 3.15 Hazard ratio forest plot of survival associated RNAs in the ceRNA network for colon, rectum, and rectosigmoid junction sites. The molecules with a hazard ratio > 1 indicate risk factors, and those with a hazard ratio < 1 indicate protective factors. Published at Vieira et al [169].

In order to further explore the patient's survival, we used the KM method to plot their survival curve. As the KM method also gives as output a list of molecules that affect the patient survival, we divided our survival curve analysis in two: for the molecules predicted by the KM method; and the molecules predicted by the KM and CoxPH methods.

Figure 3.16 shows the top two molecules from the KM method with the lowest p-value for the colon: RPS6KA5 and hsa-miR-1271-5p. We can note that in both cases the survival probability over time is diminished and that for hsa-miR-1271-5p, the patient probability of survival gets lower over time when the molecule is highly expressed.

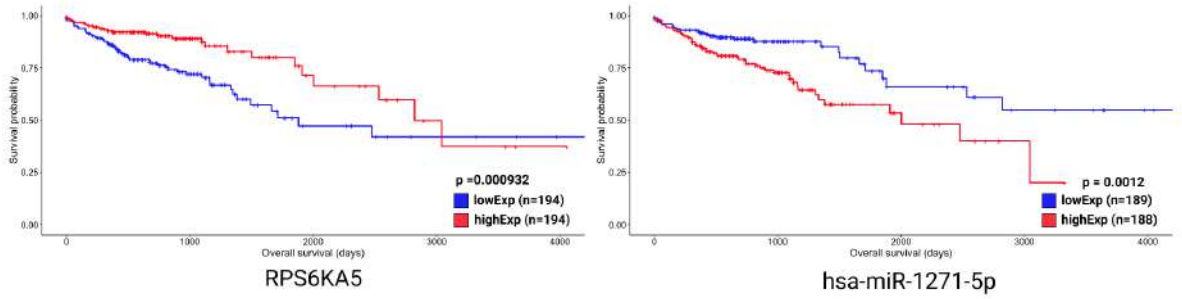

Figure 3.16 Kaplan-Meier survival curves for the two best scored molecules for the colon. Horizontal axis: overall survival time (in days), Vertical axis: survival probability.

Figure 3.17 shows the top two molecules from the KM method with the lowest p-value for rectum: E2F8 and DMXL1. We can note that in both cases the survival probability over time is diminished when the molecules are lowly expressed. Unfortunately, we can also notice that for E2F8 an event happens around 1500 days and causes the end of the survival curve, which among other reasons can be caused because the small amount patients with CRC at the rectum to provide a better analysis.

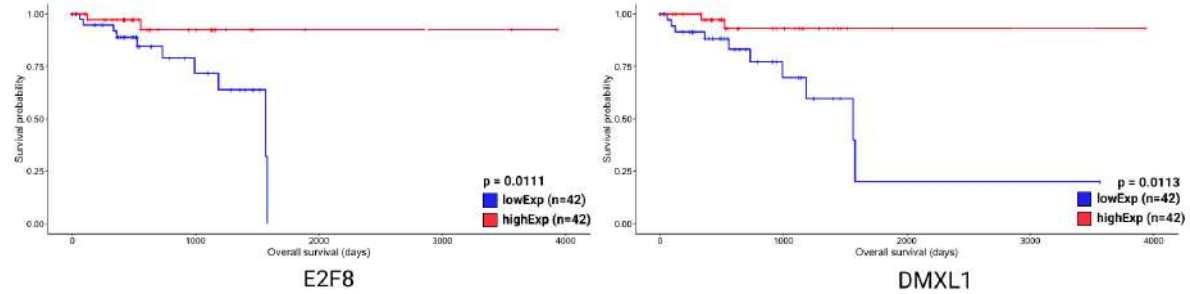

Figure 3.17 Kaplan-Meier survival curves for the two best scored molecules for the rectum. Horizontal axis: overall survival time (in days), Vertical axis: survival probability.

Figure 3.18 shows the top two molecules from the KM method with the lowest p-value for rectosigmoid junction: hsa-miR-130b-3p and AGAP3. We can note that for AGAP3 the survival probability over time is diminished when the molecules are lowly expressed and the other way around for hsa-miR-130b-3p.

After the analysis with KM, we found some PC that was not in the group of 20 molecules found with CoxPH (Figure 3.15), but that could also be relevant for overall

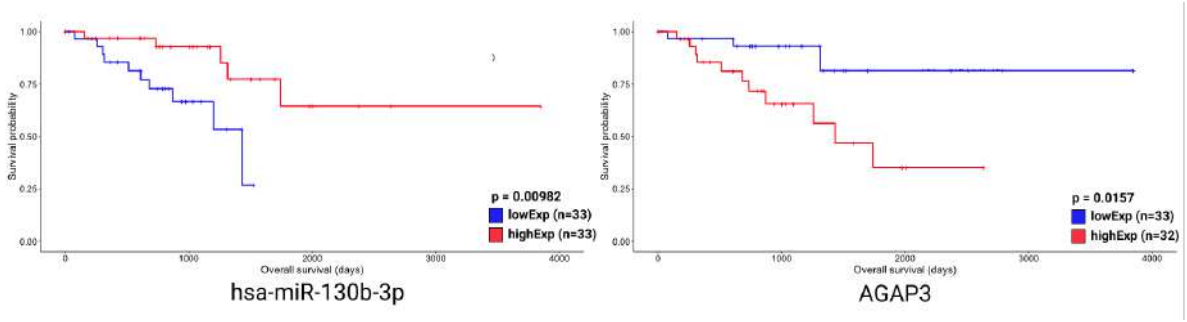

Figure 3.18 Kaplan-Meier survival curves for the two best scored molecules for the rectosigmoid junction. Horizontal axis: overall survival time (in days), Vertical axis: survival probability.

patient survival, such as: *RPS6KA5* for colon; *DMXL1* for rectum; and *AGAP3* for the rectosigmoid junction. Finally, Figure 3.19 shows the survival curves for molecules relevant both in KM and CoxPH methods. As an intersection of both methods, we have eight molecules that could be considered potential biomarkers for CRC prognosis in each anatomical site.

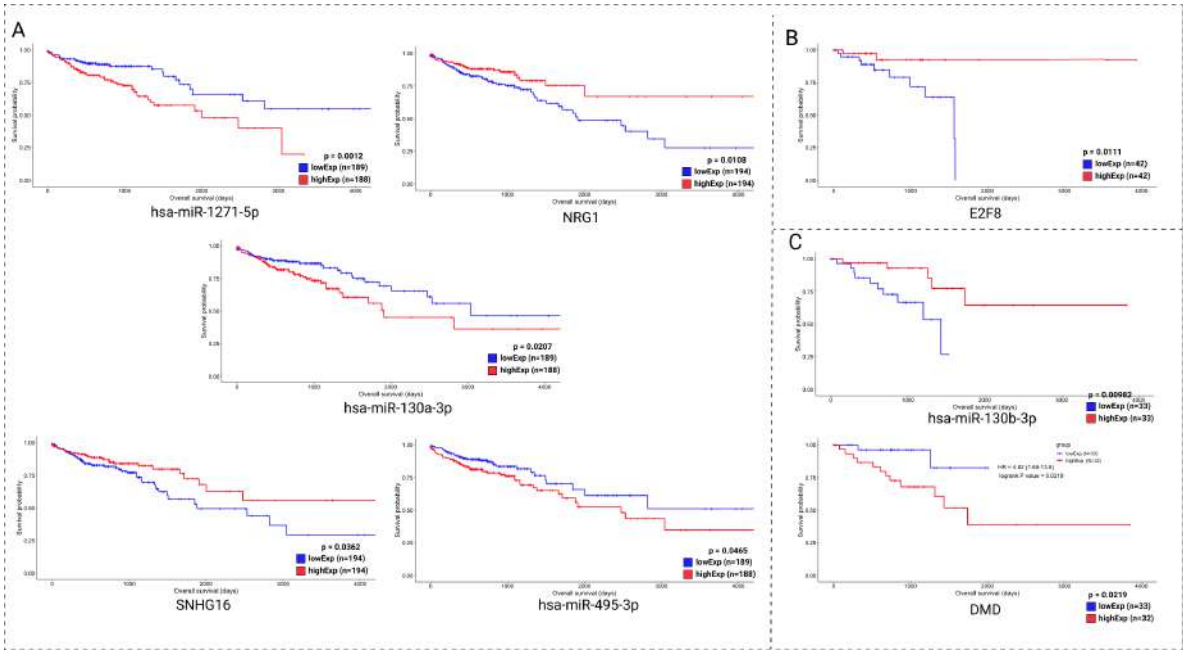

Figure 3.19 Kaplan-Meier survival curves for the best scored molecules with top HR from CoxPH for colon **(A)**, rectum **(B)**, and rectosigmoid junction **(C)** sites. Published at Vieira et al [169]

3.3 Discussion

In this study, we built a pipeline that used different bioinformatics approaches to predict potential biological markers that affect CRC prognosis. As said before, CRC is one of the most common and lethal types of cancer in Brazil and worldwide. The focus of this work was to analyze CRC occurrence in three anatomical sites: the colon, rectum, and rectosigmoid junction. It is important to perform this analysis because the chosen treatment therapy is directly related to the tumor location and the wrong diagnosis can lead to over or under-treatment. Therefore, identifying molecular markers that could help identify tumor sites and molecular characteristics is necessary. In this sense, the construction of ceRNA networks allowed us to evaluate miRNA-lncRNA-PC interactions, CRC control mechanisms, and the overall survival of patients.

Each of the steps of the pipeline gave as output potential biomarkers for CRC prognosis, but by not only performing a DE analysis in lncRNAs, PCs, and miRNAs, but also by building the ceRNA networks for each anatomical site and by performing a functional and survival analysis we further assessed the relevance of each of these molecules and their impact on patient prognosis. Also, we could further explore the unique and common factors of the CRC mechanism in each anatomical site.

When analyzing the common factors of the ceRNA networks, we expected to identify the lncRNA H19 as a protagonist, as previous studies [2, 7, 8] highlighted its protagonist in regulating CRC. However, in our analysis, H19 was present only in the colon exclusive ceRNA network. Although, in this network, H19 was pointed as a risk factor and acted as ceRNA for SOX12, ANKRD6, STC1 and hsa-miR-130a-3p, all of which are present as putative risk factors.

In the end, the common ceRNA network in colon, rectum, and rectosigmoid junction ceRNA networks was composed of four subnetworks regulated by the lncRNAs: *MAGI2-AS3*, *HAGLR-AS3*, *SNHG1* and *SNHG15*, which may suggest that these molecules play a role in CRC independent of the anatomical site (Figure 3.9). Further exploring these common networks, we can notice that the common mechanisms are related to the regulation of Wnt signaling, cell morphogenesis, and proliferation (Figure 3.14), which are known to be present in cancer. These molecules were also previously related to other known cancer pathways: MAGI2-AS3 with cell apoptosis and proliferation in CRC [177]; *HAGLR-AS3* with cell proliferation, invasion and apoptosis [186]; *SNHG1* with cell growth and promotion of CRC through the Wnt/β-catenin signaling pathway [180, 181]; and *SNHG15* with cell proliferation, apoptosis, and activation of the Wnt/β-catenin signal in CRC [178, 179]. Although these molecules were previously pointed out as related to cancer development, our study was the first to relate all of them together as common factors in CRC underlying

mechanisms and to indicate their joint use as potential biomarkers for colon, rectum, and rectosigmoid junction cancer common behavior.

Within the *MAGI2-AS3* network, we found the dystrophin gene (*DMD*). *DMD* plays a special role in muscle fiber integrity [187] and it was the only gene identified as a potentially significant risk factor in both rectum and rectosigmoid junction sites. Duchenne muscular dystrophy is a disease known to be associated with *DMD* and our functional analysis relates the biological disease's pathways from DO to the rectum ceRNA (Figure 3.11). This gene is part of a network where it is regulated by miRNAs hsa-miR-374a-5p and hsa-miR-374b-5p, and the lncRNA *MAGI2-AS3*. These three ncRNAs connected to DMD are also 'sponged' by the PC FOXO1, which is critical to tumor suppression and apoptosis [188] and presented a putative protective role in colon CRC tumors. Although Zhong et al [2] previously reported their interaction, the authors did not mention the *DMD* and FOXO1 genes, nor did they evaluate their putative role as biomarkers or as survival factors. Therefore, to the best of our knowledge, this is the first time that *DMD* is reported as a potential biomarker for poor prognosis in CRC.

In the case of the rectosigmoid junction, we found *DMD* and hsa-miR-130b-3p to be relevant to the patient prognosis. Some studies have reported the importance of hsa-miR-130b-3p in poor prognosis of CRC [189, 190]. It is worth noting that hsa-miR-130b-3p, which is relevant to the rectosigmoid junction is in the same ceRNA network as hsa-miR-130a, which is relevant to colon prognosis. Both molecules are regulated by the lncRNA MIR17HG, which may indicate that this ceRNA network is relevant to both the colon and rectosigmoid junction. However, the miRNA responsible for poor patient prognosis is different for each site.

The specific networks for the colon and rectum present distinct enriched biological pathways, with more specific endothelial development in the colon and cell morphology in the rectum. Due to the low number of samples for the rectosigmoid junction, we were unable to find a statistically significant pathway for this network. However, the pathways found are related to phosphorylation and signal transduction. These different biological pathways highlight differences in CRC behavior between distinct anatomical sites, thus reinforcing the importance of correctly identifying the tumor site.

E2F8 and *RFWD3* presented putative protective roles for rectum CRC tumors. *E2F8* encodes transcription factors that regulate development by the cell cycle [191] and *RFWD3* is known to be essential in the process of repairing DNA interstrand cross-links [192]. Both genes are connected with the lncRNA *SNHG1* but are regulated by different miRNA. The *SNHG1* ceRNA network is common for all CRC sites, but only interacts with *RFWD3* and *E2F8* in the rectum, indicating a potential role for this network in rectum cancer. The *E2F8* gene has been reported relevant to CRC as well as in regulating cancer progres-

sion [191, 193] and our survival analysis indicates better survival for high *E2F8* expression levels. Previous studies [191, 193], have identified *E2F8* as a biomarker for colon cancer, but they did not evaluate the potential role of *SNHG1-RFWD3-E2F8* ceRNA network interaction in rectum cancer.

The *RPS6KA5* gene encodes for a tyrosine kinase and has been indicated as a biomarker for colon cancer [194] through interaction with hsa-miR-130a [195]. In our colon specific network, the lncRNA MIR17HG sponges hsa-miR-130a and interacts with *RPS6KA5*. Hsa-miR-1271-5p, hsa-miR-130a, *SOX12*, *ANKRD6*, *TMEM198*, *STC1*, *H19* and *NRG1* all presented potential risk factors for colon cancer. Most of these molecules are present in distinct regions of the ceRNA network, with the exception of miR-1275-5p and *NRG1*. Both of these molecules are connected to the lncRNA MALAT1 and present opposing putative roles. Some studies [2, 7, 8] have previously reported the effects of *H19* ceRNA on CRC, but both our network and survival analyses suggest its influence only in the case of tumors located in the colon. No enrichment pathway of the rectosigmoid junction presented an exclusive HR relevant molecule.

In further consideration of the overall survival evidence, we reaffirm the potential role as prognosis biomarkers for: hsa-miR-1271-5p, *NRG1*, hsa-miR-130a-3p, *SNHG16*, and hsa-miR-495-3p, in the colon; *E2F8*, in the rectum; and of *DMD* and hsa-miR-130b-3p, in the rectosigmoid junction.

This study had some limitations. Initially, although several novel lncRNAs, *PCs* and miRNAs with clinical significance for CRC were found, the study was performed with TCGA data and no further experimental validation was carried out. Secondly, less information was analyzed for the rectum and rectosigmoid junction tissues than for the colon, which could influence site-specific results. Research on ceRNAs in CRC is still in development and requires further experimental studies and a greater amount of data from colon, rectum, and rectosigmoid cancer in order to improve our understanding of the biomarkers found.

In conclusion, this study provided a pipeline to identify potential markers that affect the patient's overall survival and underlying mechanisms for colon, rectum, and rectosigmoid junction cancer. As a byproduct of the pipeline, we construct ceRNA networks, providing clinical significance and functional implications for cancer at each of these sites. Finally, we highlighted several potential prognostic markers for CRC, and also ceRNAs that can help to explain the differences between and common factors on prognosis for the CRC sites.

Chapter 4

A biological and clinical feature analysis to predict recurrence and patient survival for CRC

In this chapter, I present a method based on ML techniques to predict CRC recurrence and patient survival. Section 4.1 contains a description of the method, along with an analysis of the biological and clinical features relevant to the construction of the prediction model. In Section 4.2 and Section 4.3 I discuss the results obtained.

4.1 A method to predict CRC recurrence and patient survival

In this section, I present the method and the data used as input, then describe the biological and clinical features extracted from the input data, and lastly, outline the generic pipeline to predict patient survival and CRC recurrence.

4.1.1 Method description and input data

The model is designed to predict two CRC patient prognosis metrics: recurrence, which indicates whether the CRC tumor grows back after treatment; and patient survival, which indicates whether a patient survives after treatment until the last known medical appointment. The generic pipeline proposed to predict CRC recurrence and patient survival, shown in Figure 4.1, is composed of two main phases: (i) data pre-processing, in which the patient's clinical and biological data is processed; and (ii) model construction, in which the prediction model is constructed and evaluated. I implemented the pipeline in Python using the scikit-learn [196] package for the ML algorithms implementation.

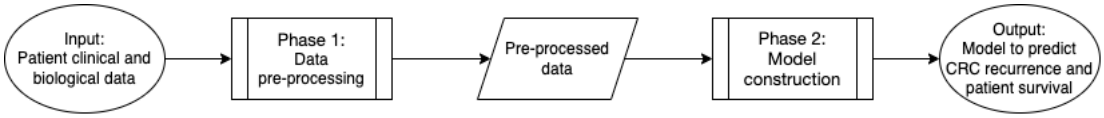

Figure 4.1 A method to predict CRC recurrence and patient survival. The pipeline is divided into two main phases: data pre-processing, in which the patient's clinical and biological data is processed; and model construction, in which the prediction model is constructed and evaluated.

As input for the method, I extracted biological and clinical information from two databases, TCGA rectal adenocarcinoma (TCGA-COAD)[1]; and TCGA rectal adenocarcinoma (TCGA-READ)[2]. I selected data exclusively from adenocarcinoma, as it is the most common for CRC, and filtered data to minimize variance, by removing possible outlier cases. Therefore, I collected RNA expression raw count data from both projects from 541 primary tumor (TP) and 48 non-tumor tissues (NT) from 545 patients, where 391 patients had colon cancer, 85 had rectum cancer and 69 had rectosigmoid junction cancer. Patient age ranged from 31 to 90 years old, with an average age of 66 years old. Of these, 185 patients (34%) received chemotherapy, 105 (19%) had a relapse, and 108 patients (20%) died. See details of the data in the GitHub of this project[3]. Next, I detail these pipeline phases.

4.1.2 Phase 1: data pre-processing

The data pre-processing phase (Figure 4.2) consists of three steps: (i) *feature extraction*, in which the clinical and biological features are extracted from the input data; (ii) *normalization*, in which the clinical and biological data are normalized to numerical values; and (iii) *missing features handler*, consisting in the creation of cases to be analyzed, according to the missing features in the data.

The feature extraction step uses the input data described in Section 4.1.1 and maps the biological and clinical features for each patient. For biological features, as proposed in Vieira et al. [169], I extracted the target biomarkers, in which: (i) the molecules are differentially expressed (DE); (ii) the biomarkers are present in the CRC ceRNA networks; and (iii) the biomarkers affect patient survival. These criteria guaranteed the selection of molecules with a potential role in the CRC patient prognosis [169] and led to the compilation of a list of nineteen molecules, as shown in Table 4.1.

In order to parameterize these molecules as biological features, I built customized R and Python scripts, to extract two key items from the RNA raw expression count data:

[1]https://portal.gdc.cancer.gov/projects/TCGA-COAD
[2]https://portal.gdc.cancer.gov/projects/TCGA-READ
[3]https://github.com/lmacielvieira/crc_pipeline

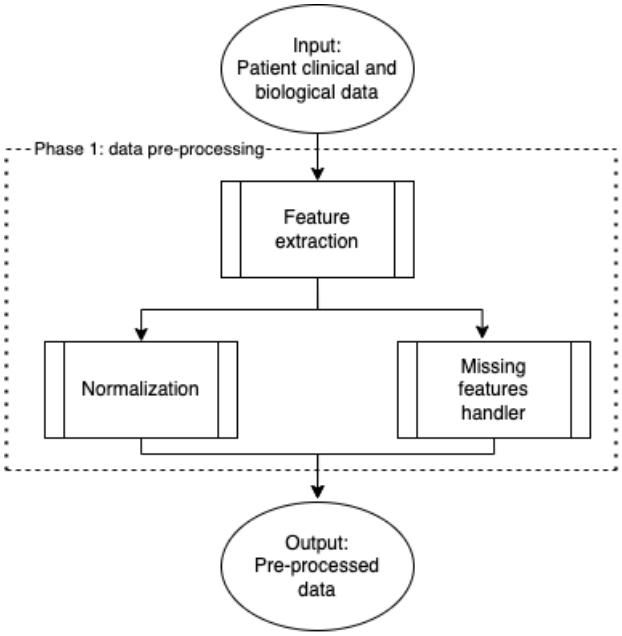

Figure 4.2 Data pre-processing phase of the method to predict CRC recurrence and patient survival. This phase receives the patient's clinical and biological data as input and consists of three steps: *feature extraction*, in which clinical and biological features from the input data; *normalization*, in which the clinical and biological data are normalized to numerical values; and *missing handler*, consisting in the creation of cases to be analyzed, according to the missing features in the data.

the molecule average expression; and the molecule expression level for each patient. To obtain the average molecule expression, I performed a DE analysis, using GDCRNATools v1.6 [170] and the Voom normalization and limma methods [171]. Then, with the average expression of each molecule and the expression level for each patient, the script verifies if each of these molecules was over-expressed in patients.

To extract the clinical features, first, I analyzed the raw clinical metadata available at TCGA. These clinical features are divided into nine groups - clinical, demographic, diagnosis, exposure, family history, follow-up, molecular test, pathology detail, and treatment. Professor João Batista de Sousa, an expert in CRC, assisted in the process of manually choosing the most relevant characteristics from the available data. The following features were chosen: age at initial pathological diagnosis; ethnicity; gender; race; vital status; number of positive lymph nodes; number of lymph nodes; pathological stage; weight; height; chemotherapy; new tumor event; and vital status.

To normalize and prepare the data to be used in the prediction models, in the *normalization* step, the clinical and biological features were transformed into numerical values, as shown in Table 4.2. These numerical values were later used in the charts that illustrate the feature importance.

Table 4.1 Candidate molecules to be used as biological features in the ML model to predict CRC recurrence

Molecule	Type	Potential roles in CRC
AGAP3	PC	cell proliferation [197]
ANKRD6	PC	immune invasion [198]
DMD	PC	lymph node metastasis [199]
E2F8	PC	cell proliferation [193]
FOXO1	PC	chemoresistance [200]
NRG1	PC	tumorigenesis [201]
SOX12	PC	cell proliferation [202]
STC1	PC	cell migration [203]
TMEM198	PC	CRC prognosis [169]
UST	PC	CRC prognosis [169]
hsa-miR-1271-5p	miRNA	cell proliferation [204]
hsa-miR-130a-3p	miRNA	cell proliferation [205]
hsa-miR-130b-3p	miRNA	cell growth [206]
hsa-miR-495-3p	miRNA	cell proliferation [207]
KCNQ1OT1	lncRNA	chemo resistence [208]
H19	lncRNA	cell migration and invasion [209]
MYLK	lncRNA	cell migration [210]
SNHG16	lncRNA	cell growth [211]
SNHG20	lncRNA	cell apoptosis [212]

Finally, in the *missing features handler* step, the data points with any missing feature were removed or associated with a value generated according to their distribution. In Section 4.2.1 I explain the results of these distinct strategies.

4.1.3 Phase 2: model construction

The model construction phase (Figure 4.3) consists of five steps: (i) *data split*, which divides the pre-processed data into train and test data; (ii) *feature selection*, in which RFE combined with LASSO and RF combined with RF select the most relevant features; (iii)*parameter optimization*, where grid search and cross-validation optimize the ML hyperparameters; and (iv) *ML classifiers construction*, in which ML models using different ML algorithms are built; and (v) *performance evaluation*, in which the ML classifiers are evaluated and compared.

First, in the *data split* step, I divided pre-processed data in a 80% and 20% ratio, to be used for training and testing, respectively. The training data then undergoes the *feature selection* step, using RFE, LASSO, and RF to select the most important features to build the ML prediction models. To interpret the features and compare feature selection behavior, I compared the RFE and LASSO combination to the RFE and RF combination.

Table 4.2 List of numerical values used in the feature vector.

Feature	Meaning	Associated Values
Age	Age of the patient	numerical value = age of the patient
Chemotherapy	If patient received chemotherapy	1 = received chemo; 0 = did not receive chemo
Ethnicity	Ethnicity of the patient	1 = latino; 0 = non latino
Gender	Gender of the patient	0 = female; 1 = male
Height	Height of the patient	numerical value = height of the patient
Race	Race of the patient	1 = non white; 0 = white
Pathological stage	CRC pathological stage	stage IV = 3; stage III = 2; stage II = 1; stage I = 0
Vital status	Vital status of the patient	1 = dead; 0 = alive
Number of positive lymph nodes	Number of positive lymph nodes in biopsy tissue	numerical value = number of lymph nodes
Number of lymph nodes	Number of lymph nodes in biopsy tissue	numerical value = number of positive lymph nodes
Weight	weight of the patient	numerical value = weight of the patient
New tumor event	CRC recurrence	1 = new tumor; 0 = no new tumor
AGAP3	AGAP3 overexpressed in the patient	1 = overexpressed; 0 = not overexpressed
ANKRD6	ANKRD6 overexpressed in the patient	1 = overexpressed; 0 = not overexpressed
DMD	DMD overexpressed in the patient	1 = overexpressed; 0 = not overexpressed
E2F8	E2F8 overexpressed in the patient	1 = overexpressed; 0 = not overexpressed
FOXO1	FOXO1 overexpressed in the patient	1 = overexpressed; 0 = not overexpressed
NRG1	NRG1 overexpressed in the patient	1 = overexpressed; 0 = not overexpressed
SOX12	SOX12 overexpressed in the patient	1 = overexpressed; 0 = not overexpressed
STC1	STC1 overexpressed in the patient	1 = overexpressed; 0 = not overexpressed
TMEM198	TMEM198 overexpressed in the patient	1 = overexpressed; 0 = not overexpressed
UST	UST overexpressed in the patient	1 = overexpressed; 0 = not overexpressed
hsa-miR-1271-5p	hsa-miR-1271-5p overexpressed in the patient	1 = overexpressed; 0 = not overexpressed
hsa-miR-130a-3p	hsa-miR-130a-3p overexpressed in the patient	1 = overexpressed; 0 = not overexpressed
hsa-miR-130b-3p	hsa-miR-130b-3p overexpressed in the patient	1 = overexpressed; 0 = not overexpressed
hsa-miR-495-3p	hsa-miR-495-3p overexpressed in the patient	1 = overexpressed; 0 = not overexpressed
KCNQ1OT1	KCNQ1OT1 overexpressed in the patient	1 = overexpressed; 0 = not overexpressed
H19	H19 overexpressed in the patient	1 = overexpressed; 0 = not overexpressed
MYLK	MYLKoverexpressed in the patient	1 = overexpressed; 0 = not overexpressed
SNHG16	SNHG16 overexpressed in the patient	1 = overexpressed; 0 = not overexpressed
SNHG20	SNHG20 overexpressed in the patient	1 = overexpressed; 0 = not overexpressed

Then, I used SHAP to visualize the impact of each of the selected features in CRC recurrence and patient survival prediction (Figure 4.4).

In the *ML classifiers construction* step, I used six classifiers to predict CRC recurrence and patient survival: Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT) and Adaptive Boosting (AB). As described in Section 2.3 of Chapter 2, each classifier uses a different approach to predict an outcome, and classifiers behave differently based on the pattern of the input data. Therefore, the goal was to explore these classifiers in order to find the best option to predict the expected outcome. Finally, in the *performance evaluation* step, I evaluated the ML model's performance with the test data as input and compared the models through several metrics, such as accuracy, precision, and recall.

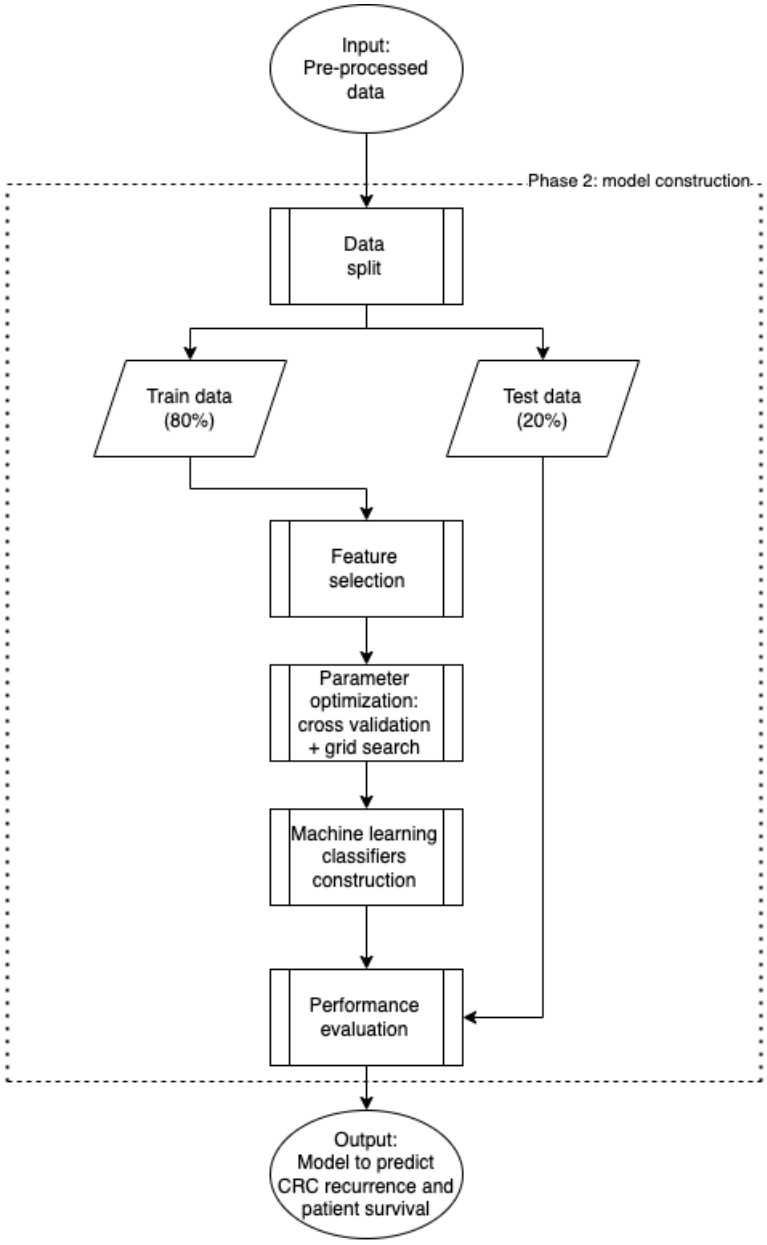

Figure 4.3 Model construction phase of the method to predict CRC recurrence and patient survival. This phase receives the pre-processed data from Phase 1 and consists of five steps: *data split*, in which the pre-processed data is divided into 80% train and 20% test; *feature selection*, in which RFE combined with LASSO and RF combined with RF select the most relevant features; *parameter optimization*, using grid search and cross-validation to optimize the ML hyperparameters; *ML classifiers construction*, building ML models using different ML algorithms; and *performance evaluation*, in which the ML classifiers are evaluated and compared.

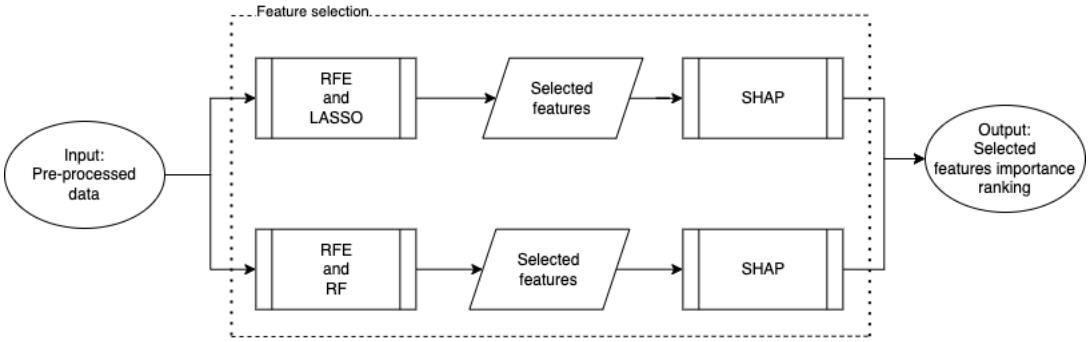

Figure 4.4 The *feature selection* step, with the pre-processed data as input, compares two different methods: RFE combined with LASSO, and RFE combined with RF. Both methods' output is a list of selected features, which are used as input to the SHAP explainer to easily visualize their individual impact on the model prediction.

4.2 Results

In this section, I present the results obtained from the proposed method, beginning with Phase 1 results - data pre-processing, followed by Phase 2 results for each prediction target - model construction.

4.2.1 Phase 1: data pre-processing

Initially, during the *feature extraction* step, I performed a DE analysis using GDCRNA-Tools v1.6 [170] and then extracted the average expression count of each DE gene. Based on the average expression, I verified whether the gene was expressed for each patient and used this as a biological feature associated with the patient. For the clinical features, I noticed that in many cases, information necessary to collect clinical metadata chosen by the specialist was missing.

Of these features, the ones with the most missing values were race, ethnicity, weight, and height. To address this problem, first, I divided the clinical features into two groups: (1) age at initial pathological diagnosis, gender, number of lymph nodes, number of positive lymph nodes, chemotherapy, pathologic stage, vital status, and new tumor event; and (2) race, ethnicity, weight, and height.

Then, considering these two groups of clinical features, I grouped data into three cases in the *missing features handler* step:

1. Filtered data with missing biological or group (1) clinical features;
2. Filtered data with missing biological or group (1) or group (2) clinical features; and
3. all data, but replacing missing clinical features by using the most frequent value. In this case, I chose the most frequent value because features like race are fixed values, and other missing data replacement techniques, like mean and median, could generate non-existing features.

After filtering and transforming the data for each case, as described, the number of patients was: for case (1), 357 with colon cancer, 74 with rectum cancer, and 63 with rectosigmoid junction cancer; for case (2), 177 with colon cancer, 27 with rectum cancer and 33 with rectosigmoid junction cancer; and for case (3), 391 with colon cancer, 85 with rectum cancer, and 69 with rectosigmoid junction cancer. With all the features set up, I proceeded to construct the prediction models for cases (1), (2), and (3).

4.2.2 Phase 2: model construction

Patient survival

In the first step of the *model construction* phase, after dividing data for training and testing, I compared two approaches (RFE combined with LASSO, and RFE combined with RF) in the *feature selection* step. Table 4.3 shows the features selected for cases (1), (2), and (3) for the patient survival prediction.

Table 4.3 List of features selected to predict patient survival, according to each designed case.

Feature	RFE + LASSO			RFE + RF[4]		
	Used in case (1)	**Used in case (2)**	**Used in case (3)**	**Used in case (1)**	**Used in case (2)**	**Used in case (3)**
Age	Yes	Yes	Yes	Yes	Yes	Yes
Positive lymph node count	Yes	Yes	Yes	Yes	Yes	Yes
Lymph node count	Yes	Yes	Yes	Yes	Yes	Yes
Pathological stage	Yes	Yes	Yes	Yes	Yes	Yes
Recurrence	Yes	Yes	Yes	Yes	Yes	Yes
Chemotherapy	Yes	Yes	Yes	Yes	Yes	Yes
hsa-miR-130b-3p	Yes	Yes	Yes	Yes	Yes	No
hsa-miR-495-3p	Yes	Yes	Yes	Yes	Yes	No
KCNQ1OT1	Yes	Yes	Yes	Yes	Yes	No
SNHG16	Yes	Yes	No	Yes	Yes	No
SNHG20	Yes	Yes	No	Yes	Yes	No
SOX12	Yes	Yes	No	Yes	Yes	No
STC1	Yes	Yes	No	Yes	Yes	No
TMEM198	Yes	Yes	No	Yes	Yes	No
Gender	Yes	Yes	No	Yes	Yes	No
Weight	No	Yes	Yes	No	Yes	Yes
Height	No	Yes	Yes	No	Yes	Yes
MYLK	Yes	Yes	No	No	Yes	No
NRG1	Yes	Yes	No	No	Yes	No
AGAP3	Yes	No	No	Yes	Yes	No
hsa-miR-130a-3p	Yes	Yes	No	No	Yes	No
Race	No	Yes	Yes	No	Yes	No
E2F8	Yes	No	No	Yes	Yes	No
ANKRD6	Yes	No	No	No	Yes	No
DMD	Yes	No	No	No	Yes	No
Ethnicity	No	Yes	No	No	Yes	No
FOXO1	Yes	No	No	Yes	No	No
H19	No	No	No	No	Yes	Yes
hsa-miR-1271-5p	No	No	No	No	Yes	No
UST	No	Yes	No	No	No	No

[4]Only those features selected to train the best ML models to predict patient survival are portrayed, in contrast to the LASSO approach, the features selected by RF may change according to the constructed ML model.

Note, in Table 4.3, that according to the RFE approach using LASSO and RF, for all three cases, clinical and biological features were relevant as input for the models. There are many similarities in the features chosen using LASSO and RF. Specifically, for both selection algorithms, six clinical features were selected for the cases: age; lymph node count; positive lymph node count; pathological stage; recurrence; and chemotherapy. It is also worth noting that LASSO seems to select more biological features than RF and that for both LASSO and RF, we have some biological features that seem to be important for cases (1) and (2): hsa-miR-130b-3p; hsa-miR-495-3p; KCNQ1OT1; SNHG16; SNHG20; SOX12; STC1; TMEM198. Figure 4.5 shows the impact of each chosen feature on the model prediction in detail, using the SHAP explainer for Case 1.

Figure 4.5 (i) shows that when using LASSO, many of the chosen features have an average impact near zero, and only seven features seem to have an overall impact in the final prediction. Through the RF approach (Figure 4.5(ii)), the importance of selected features is more distributed. Both approaches highlight the potential importance of the biological marker E2F8, and the higher the pathological stage, age, and CRC recurrence, the lower the chance of patient survival. Also, chemotherapy treatment seems to increase patient survival. Figure 4.6 shows the impact of each chosen feature on the model prediction in detail, using the SHAP explainer for Case 2.

Figure 4.6 (i) shows as in Case 1 for LASSO, many of the chosen features have an average impact near zero, and that only seven features seem to have an overall impact in the final prediction. Through the RF approach (Figure 4.6 (ii)), the importance of the selected features is more distributed. Unlike Case 1, LASSO indicated less importance of the biological marker E2F8 as compared to the newly added clinical features, *weight* and *height*. The RF approach indicates greater importance to biological features than the LASSO approach, and, as in Case 1, points out the importance of E2F8. The observations on the pathological stage, age, CRC recurrence, and chemotherapy are also confirmed in this case. Figure 4.7 shows the impact of each feature on the model prediction in detail, using the SHAP explainer for Case 3.

Figure 4.7 (i) shows that, unlike Cases 1 and 2, in Case 3, which has more data and uses all the features, the feature importance is more distributed for both approaches. The weight and height are not indicated as important in comparison to Case 2. The RF approach does not show biological features to have a relevant impact on the final prediction, while the LASSO approach highlights the biomarkers KCNQ1OT1, has-miR-495-3p, and hsa-miR-130b-3p. The observations related to pathological stage, age, CRC recurrence, and chemotherapy are also confirmed in this case. Lastly, although Case 3 worked with some generated values (completing any missing clinical feature patient data with the most frequent value), this approach reduced the number of chosen features and

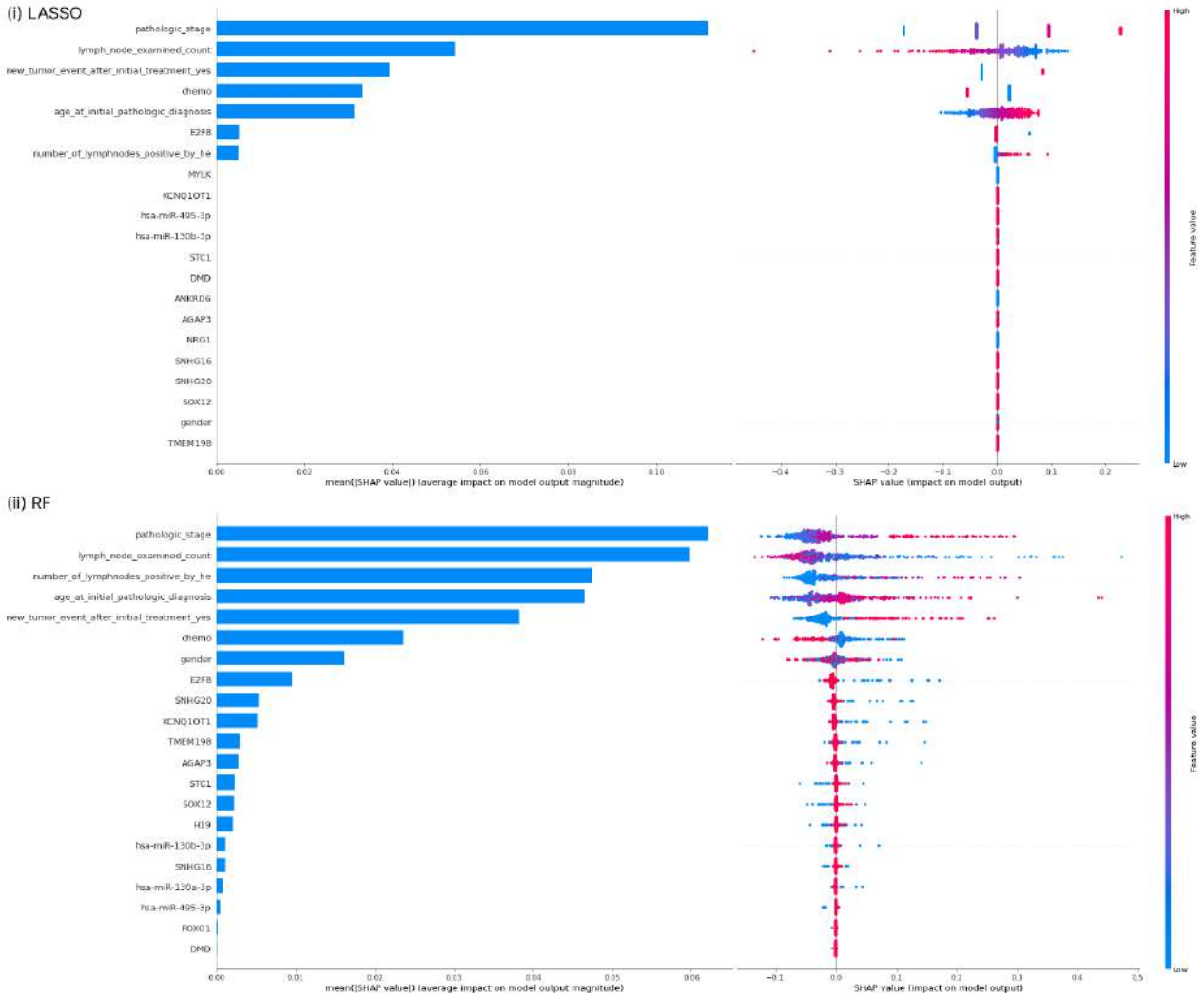

Figure 4.5 SHAP summary plot showing the importance of the features selected using RFE with (i) LASSO and (ii) RF to predict patient survival for Case 1. A positive SHAP value indicates a positive impact on prediction, leading the model to predict 1 (Patient died) and a negative value indicates a negative impact, leading the model to predict 0 (Patient survived). The bar plots on the left show the average impact of the feature in the model. The scatter plot, on the right side, is depicted such that each point on the chart is one SHAP value for a prediction and a feature, red indicating the higher value of a feature and blue indicating the lower value of a feature. The chart illustrates, for example, for both (i) and (ii), that the higher the pathological stage value, the higher the chance of patient fatality.

had a better distribution of feature impact in the prediction of patient survival.

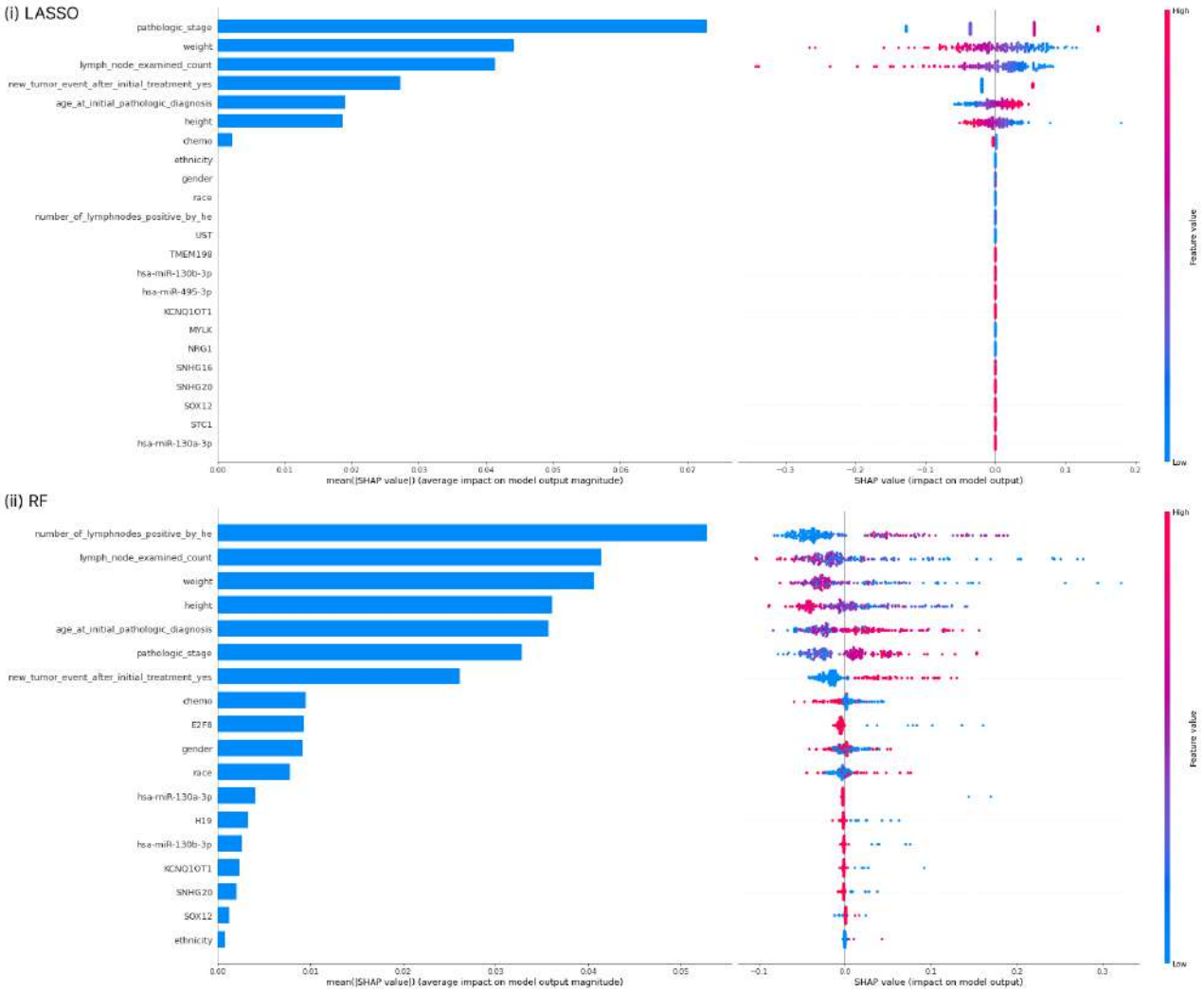

Figure 4.6 SHAP summary plot showing the importance of the features selected using RFE with (i) LASSO and (ii) RF to predict patient survival for Case 2. A positive SHAP value indicates a positive impact on prediction, leading the model to predict 1 (Patient died) and a negative value means negative impact, leading the model to predict 0 (Patient survived). The bar plots on the left show the average impact of each feature on the model. The scatter plot, on the right, is depicted such that each point on the chart is one SHAP value for a prediction and a feature, red indicating the higher value of a feature and blue indicating the lower value of a feature. The chart illustrates, for example, for both (i) and (ii), that the higher the pathological stage value, the higher the chance of patient fatality.

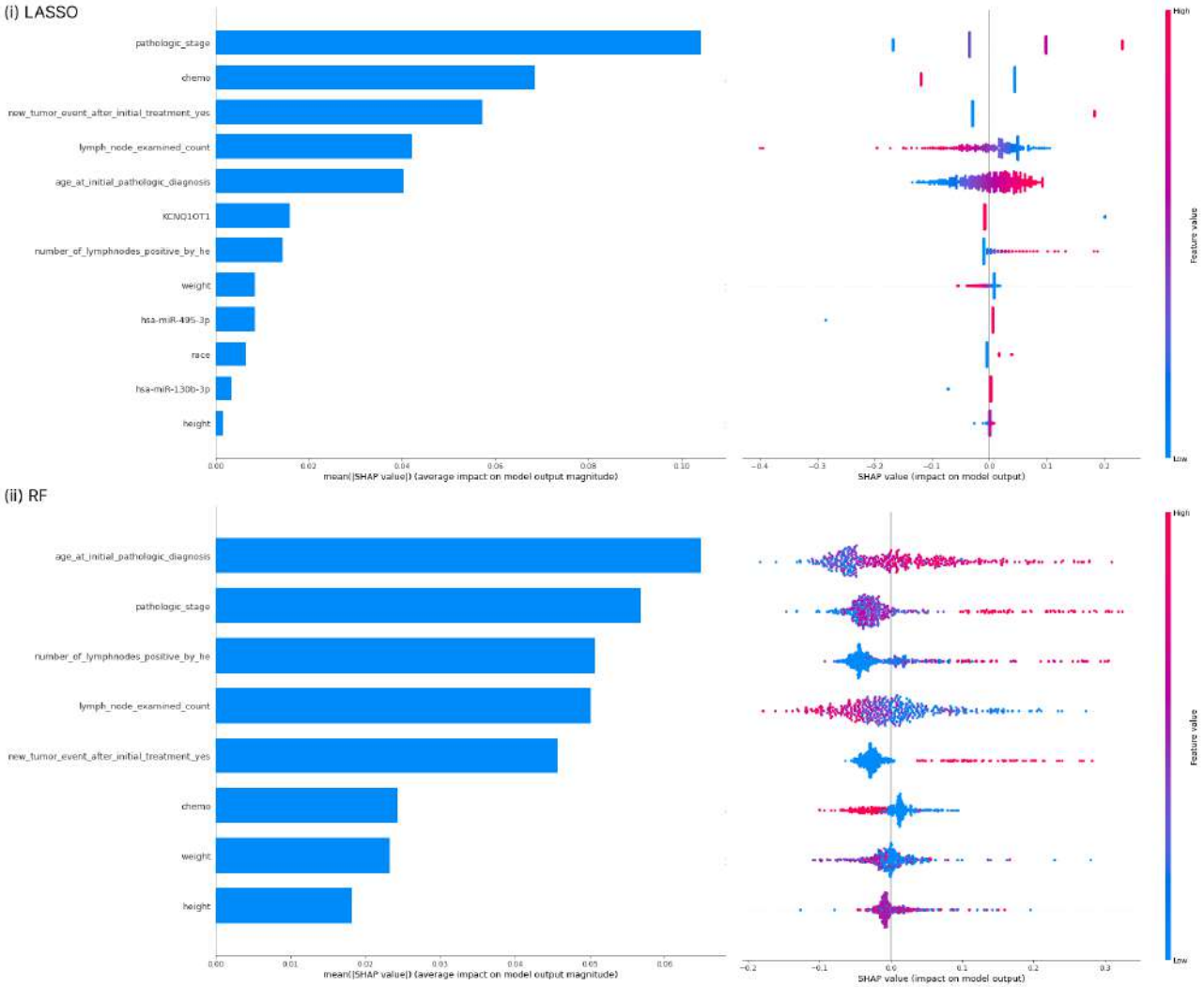

Figure 4.7 SHAP summary plot showing the importance of the features selected using RFE with (i) LASSO and (ii) RF to predict patient survival for Case 3. A positive SHAP value indicates a positive impact on prediction, leading the model to predict 1 (Patient died) and a negative value indicates negative impact, leading the model to predict 0 (Patient survived). The bar plots on the left show the average impact of the feature in the model. The scatter plot, on the right side, is depicted such that each point on the chart is one SHAP value for a prediction and a feature, red indicating the higher value of a feature and blue indicating the lower value of a feature. The chart illustrates, for example, for both (i) and (ii), that the higher the pathological stage value, the higher the chance of patient fatality.

After selecting features with RFE using RF and LASSO as described, I proceeded to train the ML models. I obtained the ML models optimized parameters using grid search, listed in the project repository[5]. Table 4.4 shows the performance evaluation of all the models constructed for predicting patient survival for all cases, using the features selected by each of the RFE approaches.

Table 4.4 Performance evaluation of the ML models, used to predict patient survival in all cases, using the features selected by each of the RFE approaches.

	RFE + LASSO			RFE + RF		
Model	**Accuracy for Case (1)**	**Accuracy for Case (2)**	**Accuracy for Case (3)**	**Accuracy for Case (1)**	**Accuracy for Case (2)**	**Accuracy for Case (3)**
SVM	80%	73%	74%	79%	73%	75%
LR	**85%**	**90%**	79%	**85%**	**90%**	79%
KNN	73%	73%	69%	73%	73%	68%
DT	75%	81%	77%	74%	73%	80%
AB	78%	85%	82%	78%	85%	80%
RF	84%	81%	**83%**	81%	81%	**83%**

I obtained the following results in predicting patient survival. For the RFE combined with LASSO approach, the LR model led to the best accuracy for Case (1), achieving an accuracy of 85% on the test data with a 78% precision and 67% recall. The LR model led to the best accuracy for Case (2), achieving an accuracy of 90% on the test data with 94% precision and 72% recall. The RF model led to the best accuracy for Case (3), achieving an accuracy of 83% on the test data with a precision of 75% and a recall of 62%. For the RFE combined with the RF approach, the LR model led to the best accuracy for Case (1), achieving an accuracy of 85% on the test data with 78% precision and 67% recall. The LR model led to the best accuracy for Case (2), achieving an accuracy of 90% on the test data with 94% precision and 72% recall. The RF model led to the best accuracy for Case (3), achieving an accuracy of 83% on the test data with 74% precision and 64% recall. As shown, the LR models displayed the best average performance in all cases, followed by RF.

Recurrence

In the first step of the *model construction* phase, after dividing data between training and testing, we proceeded to the *feature selection* step, comparing the approaches: RFE combined with LASSO; and RFE combined with RF. Table 4.5 shows the features selected for Cases (1), (2), and (3) for the prediction of CRC recurrence.

[5]https://github.com/lmacielvieira/crc_pipeline

Table 4.5 List of features selected to predict CRC recurrence, according to each designed case.

Feature	RFE + LASSO			RFE + RF[6]		
	Used in Case (1)	**Used in Case (2)**	**Used in Case (3)**	**Used in Case (1)**	**Used in Case (2)**	**Used in Case (3)**
Positive lymph node count	Yes	Yes	Yes	Yes	Yes	Yes
Lymph node count	Yes	Yes	Yes	Yes	Yes	Yes
Pathological stage	Yes	Yes	Yes	Yes	Yes	Yes
Chemotherapy	Yes	Yes	Yes	Yes	Yes	Yes
SNHG16	Yes	Yes	Yes	Yes	Yes	Yes
Age	Yes	Yes	No	Yes	Yes	Yes
hsa-miR-130b-3p	Yes	Yes	No	Yes	Yes	Yes
Gender	Yes	Yes	No	Yes	Yes	Yes
hsa-miR-495-3p	Yes	Yes	No	Yes	Yes	No
SNHG20	Yes	Yes	No	Yes	Yes	No
SOX12	Yes	Yes	No	Yes	Yes	No
AGAP3	Yes	Yes	No	Yes	Yes	No
KCNQ1OT1	No	Yes	No	Yes	Yes	Yes
STC1	No	Yes	No	Yes	Yes	No
TMEM198	No	Yes	No	Yes	Yes	No
Weight	No	Yes	No	No	Yes	Yes
Height	No	Yes	No	No	Yes	Yes
Race	No	Yes	No	No	Yes	Yes
E2F8	No	Yes	No	Yes	Yes	No
H19	No	Yes	No	Yes	Yes	No
MYLK	Yes	Yes	No	No	No	No
NRG1	Yes	Yes	No	No	No	No
hsa-miR-130a-3p	No	Yes	No	No	Yes	No
Ethnicity	No	Yes	No	No	Yes	No
ANKRD6	No	Yes	No	No	No	No
DMD	No	Yes	No	No	No	No
hsa-miR-1271-5p	No	Yes	No	No	No	No
UST	No	Yes	No	No	No	No
FOXO1	No	No	No	No	No	No

Table 4.5 illustrates that for all three cases the RFE approach using LASSO and RF selected clinical and biological features as relevant to be used as input to the models. There are many similarities in the features chosen using LASSO and RF. In particular, both selection algorithms chose the following five features for all cases: lymph node count; positive lymph node count; pathological stage; chemotherapy; and SNHG16. The LASSO approach was more conservative for Case 3 and mapped few features as relevant. Figure 4.8 shows the impact of each feature on the model prediction in detail with the SHAP explainer for Case 1.

Figure 4.8 highlights the potential importance of both biological and clinical markers

[6]Only those features selected to train the best ML models to predict CRC recurrence are portrayed, in contrast to LASSO, the features selected by RF may change according to the constructed ML model.

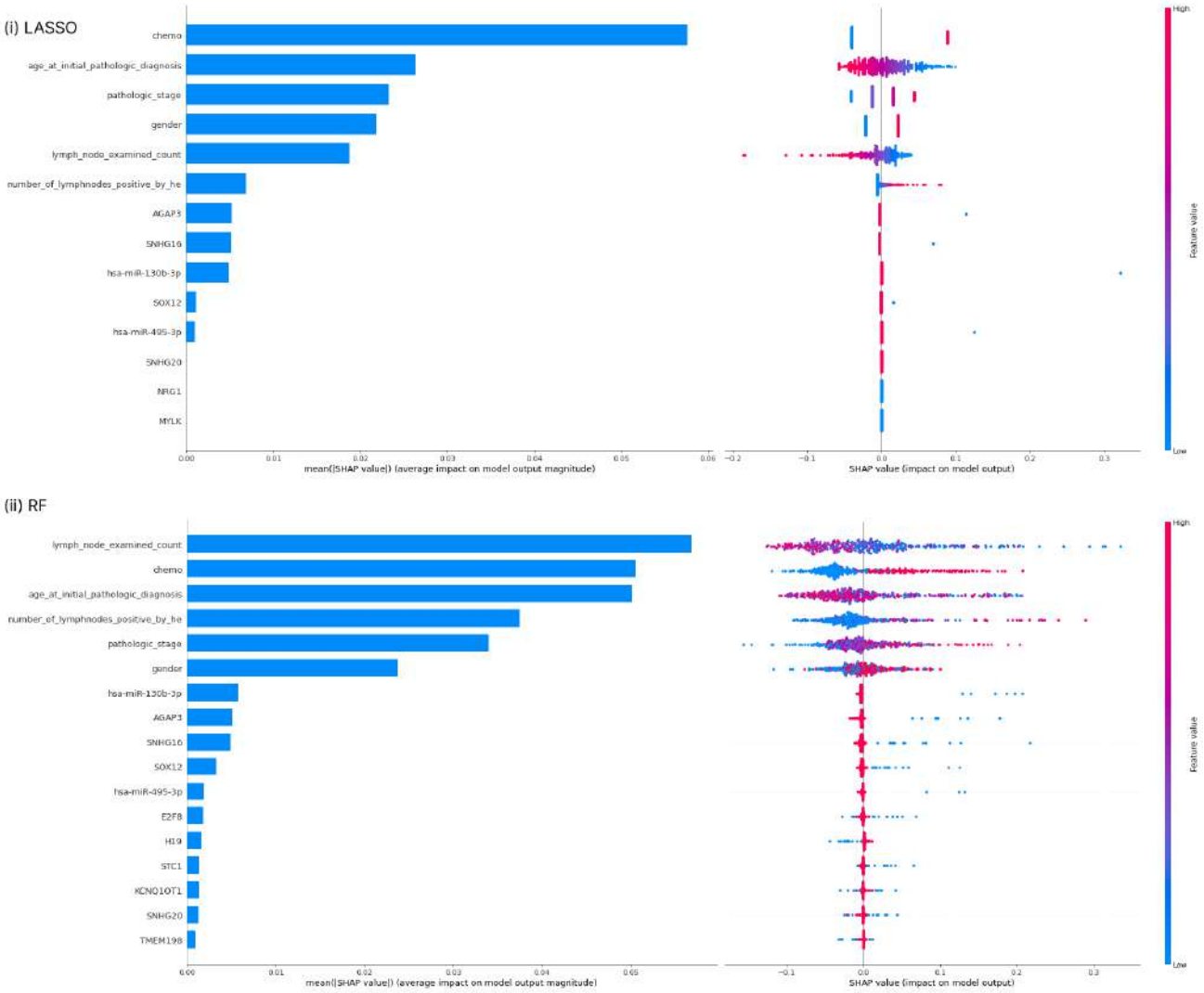

Figure 4.8 SHAP summary plot showing the importance of features selected using RFE with (i) LASSO and (ii) RF, to predict CRC recurrence for Case 1. A positive SHAP value indicates a positive impact on prediction, leading the model to predict 1 (New tumor) and a negative value indicates negative impact, leading the model to predict 0 (No tumor). The bar plots on the left show the average impact of the feature in the model. The scatter plot, on the right side, is depicted such that each point on the chart is one SHAP value for a prediction and a feature, red indicating the higher value of a feature and blue indicating the lower value of a feature. The chart shows, for example, that for both (i) and (ii), the higher the pathological stage value, the higher the chance of a new tumor event.

in the prediction of CRC recurrence. As in the analysis for patient survival, the higher the pathological stage, the higher the risk of a new tumor event. The LASSO approach indicates that male patients (gender = 1) have a higher chance of CRC recurrence. Figure 4.9 shows the impact of each feature on the model prediction in detail with the SHAP explainer for Case 2.

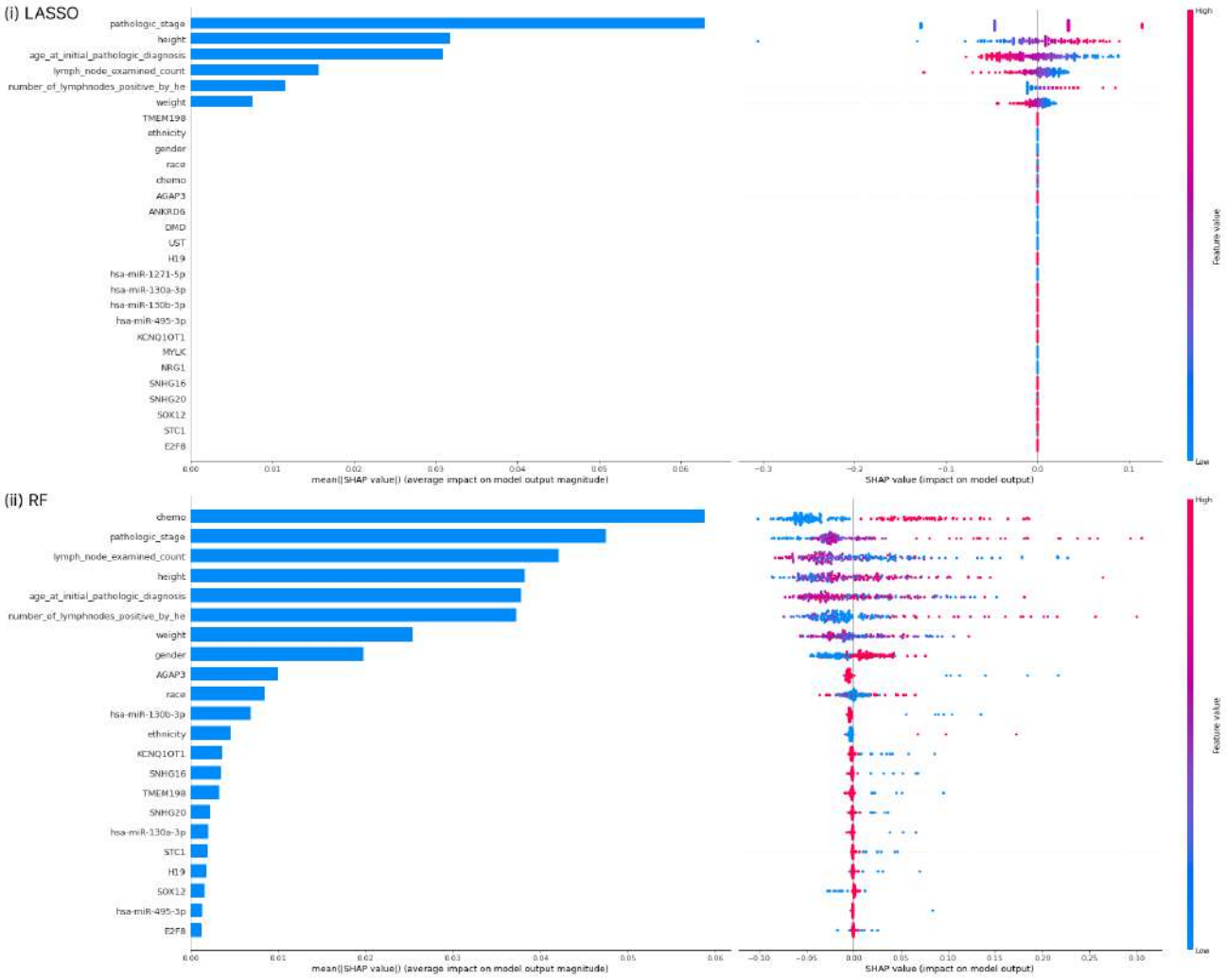

Figure 4.9 SHAP summary plot showing the importance of the features selected using RFE with (i) LASSO and (ii) RF to predict CRC recurrence for Case 2. A positive SHAP value indicates a positive impact on prediction, leading the model to predict 1 (New tumor) and a negative value indicates negative impact, leading the model to predict 0 (No tumor). The bar plots on the left show the average impact of the feature in the model. The scatter plot, on the right side, is depicted such that each point on the chart is one SHAP value for a prediction and a feature, red indicating the higher value of a feature and blue indicating the lower value of a feature. The chart shows, for example, that for both (i) and (ii), the higher the pathological stage value, the higher the chance of a new tumor event.

Figure 4.9 (i) shows that many of the features chosen through LASSO have an average impact near zero, and that only six features seem to have a high overall impact on the final prediction. Through the RF approach (Figure 4.6 (ii)) the importance of selected features is more distributed. The RF approach gives more importance to biological features as

compared to the LASSO approach. The observations related to the pathological stage are again confirmed in this case. Figure 4.10 shows the impact of each feature on the model prediction in detail with the SHAP explainer for Case 3.

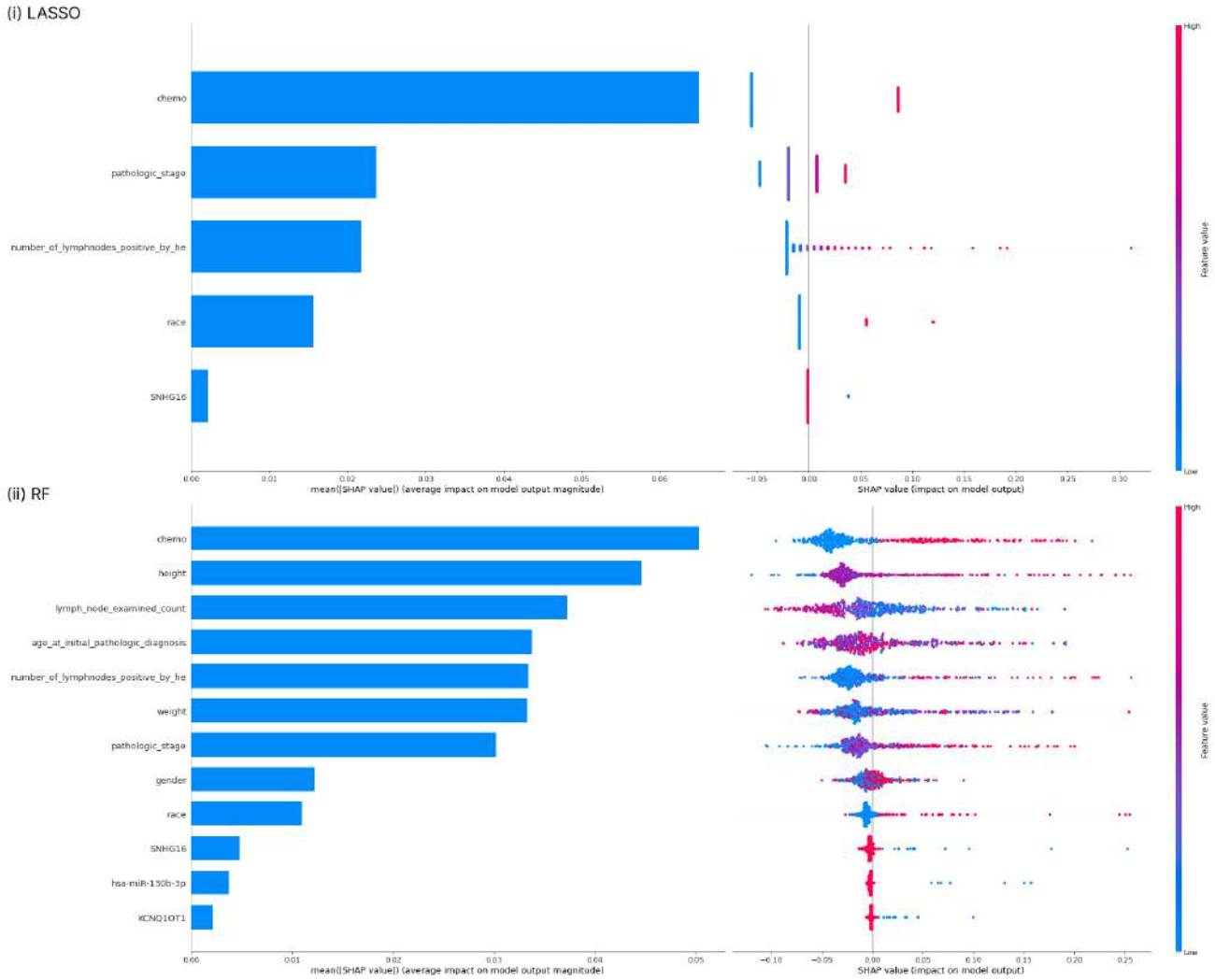

Figure 4.10 SHAP summary plot showing the importance of the features selected using RFE with (i) LASSO and (ii) RF to predict CRC recurrence for Case 3. A positive SHAP value indicates a positive impact on prediction, leading the model to predict 1 (New tumor) and a negative value indicates negative impact, leading the model to predict 0 (No tumor). The bar plots on the left show the average impact of the feature in the model. The scatter plot, on the right side, is depicted such that each point on the chart is one SHAP value for a prediction and a feature, red indicating the higher value of a feature and blue indicating the lower value of a feature. The chart illustrates, for example, that for both (i) and (ii) the higher the pathological stage value, the higher the chance of a new tumor event.

Figure 4.10 (i) shows that unlike Case 2, in Case 3, which has more data and uses all

the features, the LASSO approach leads to a greater distribution in feature importance. In this case, the RF approach contains all the features selected by LASSO. It is worth noting that the only biomarker in both cases is SNHG16. Although this approach works with some generated values (replacing any missing clinical feature patient data with the most frequent value), it also filters the number of features more and has a better distribution of feature impact in predicting patient survival.

After selecting features with RFE using RF and LASSO as described, I proceeded to train the ML model. I obtained the ML models optimized parameters using grid search and are listed in the project repository[7]. Table 4.6 shows the performance evaluation of all the models constructed for predicting CRC recurrence in all cases, using the features selected by each RFE approach.

Table 4.6 Performance evaluation of the ML models, used to predict CRC recurrence in all cases, using the features selected by each RFE approach.

	RFE + LASSO			RFE + RF		
Model	**Accuracy for Case (1)**	**Accuracy for Case (2)**	**Accuracy for Case (3)**	**Accuracy for Case (1)**	**Accuracy for Case (2)**	**Accuracy for Case (3)**
SVM	80%	75%	80%	79%	73%	80%
LR	80%	**79%**	81%	**82%**	**77%**	80%
KNN	77%	65%	**82%**	78%	62%	81%
DT	79%	75%	76%	80%	62%	80%
AB	78%	73%	78%	79%	73%	81%
RF	**82%**	75%	80%	81%	75%	**82%**

I obtained the following results for the prediction of CRC recurrence. For the approach combining RFE with LASSO, the RF model led to the best accuracy for Case (1), achieving an accuracy of 82% on the test data with 91% precision and 54% recall. The LR model led to the best accuracy for Case (2), achieving an accuracy of 79% on the test data with 74% precision and 64% recall. The KNN model led to the best accuracy for Case (3), achieving an accuracy of 82% on the test data with 70% precision and 60% recall. For the approach combining RFE with RF, the LR model led to the best accuracy for Case (1), achieving an accuracy of 82% on the test data with 91% precision and 53% recall. The LR model led to the best accuracy for Case (2), achieving an accuracy of 77% on the test data with 69% precision and 62% recall. The RF model led to the best accuracy for Case (3), achieving an accuracy of 82% on the test data with 71% precision and 57% recall. Note that the LR models displayed the best average performance in all cases, followed by RF.

[7]https://github.com/lmacielvieira/crc_pipeline

4.3 Discussion

As described in Chapter 2, Section 2.3, feature selection and ML methods are broadly used to better understanding of data as well as to generate information [120]. With the significant growth of biological data on CRC and the amount of information that can be extracted from this data for the study of CRC prognosis, the use of feature extraction techniques seems to be of interest for improving ML methods.

In this chapter, I compared feature selection methods for identifying biological and clinical features relevant to CRC recurrence and patient survival. I also proposed ML models to predict CRC recurrence and patient survival, which can help specialists to better understand key points in CRC prognosis. The proposed method combines biological and clinical features to predict CRC recurrence and patient survival, using data from patients from the United States, available in the TCGA database, as input. Using LR and RF I achieved at best accuracy of 90% and 82% for patient survival and CRC recurrence, respectively.

Previous studies [9, 10, 11] devised models to predict CRC-related outcomes through a variety of ML techniques. Gründner et al. [9] proposed a method that combines biological and clinical features to predict prognosis aspects for CRC patients from the Erlangen University Hospital. Their best model used DT to predict patient relapse (CRC recurrence) and achieved an accuracy of 71%, 73% specificity, and of 63% sensitivity. Achilonu et al. [10] created a pipeline using clinical features to predict CRC recurrence and survival in South African patients. Specifically, their best model used an artificial neural network (ANN) and achieved an accuracy of 87.0% and 82.0%, for CRC recurrence and patient survival respectively. Gupta et al. [11] described a model using clinical features to predict colon cancer stages and DFS from Chang Gung Memorial Hospital patients. Their best model used RF and achieved an AUC of 89.0% and 84.0%, for cancer stages and DFS, respectively. Table 4.7 summarizes these methods with the features used and best accuracies.

Table 4.7 Methods based on ML to predict CRC prognosis.

Method	Demographics	Biological features	Clinical features	Best accuracy
Gründner et al. [9]	Germany	list of 58 genes	Localization, gender, smoker, weight, height, cancer type, and tumor stage	71%
Achilonu et al. [10]	South Africa	None	Race, histology, recurrence, radiological stage, language prior CRC treatment, hospital, and CRC related complication	87%
Gupta et al. [11]	Taiwan	None	Age, gender, hypertension, diabetes, smoker, alcohol, family history, and body mass index	89%
LR model (this work)	USA	hsa-miR-130b-3p, hsa-miR-495-3p, and KCNQ1OT1	Age, weight, height, chemotherapy, CRC recurrence, pathological stage, race, count and number of positive lymph nodes	90%

It is of note that the methods described in these studies, including the present study, used different data and features as input, and present relatively good accuracy in predicting specific patient prognosis factors. As shown in Table 4.7, ML methods use various clinical features to predict CRC prognosis targets. Gründner et al. [9] and the present study are the only ones to propose a method that combines biological and clinical features, and this study achieves a higher best accuracy. Achilonu et al. [10] and Gupta et al. [11] show that clinical features, such as age, gender, race, recurrence, chemotherapy, smoking, and alcohol consumption, can also lead to good accuracy in predicting CRC prognosis factors. This study identified some of the clinical features in the cited works as relevant, such as age, gender, race, recurrence, and chemotherapy. Although smoking and alcohol consumption have been shown to be relevant in the cited works [10, 11, 9]. These clinical features were not included in this study, because their values were missing from the available TCGA data for most patients. Finally, this study demonstrates that even when some relevant clinical features, like smoking and alcohol consumption are excluded, the combination with biological features seems to maintain prediction accuracy.

Other than the similarities among the chosen features and the fact that each study indicated a different ML classifier as the best predictor, most studies reported good results with RF and LR predictors, which was confirmed in the present study. Gründner et al. [9] reported low sensitivity values of their prediction models, which was confirmed in this thesis. The best algorithms and the performance evaluation patterns for sensitivity suggest a possible pattern in prediction behavior with CRC data, even with data gathered from different sources.

This study can also be used as support in planning patient treatment by providing more information for CRC prognosis. Furthermore, this study demonstrates that biological markers help to predict patient prognosis. The biological features with greatest average impact in all cases: SNHG16, hsa-miR-130b-3p, hsa-miR-495-3p and KCNQ1OT1 were also pointed out by other studies [169, 213, 214, 207, 189, 190, 215, 216] as important in the development of CRC. The results of this thesis also show that age, ethnicity, pathological stage, chemotherapy, and lymph node count, clinical features confirmed to be relevant though previous studies [217, 218, 219, 220, 221], are important even when combined with biological features. The systematic analysis comparing RFE combined with LASSO and RFE combined with RF showed that both algorithms behave similarly since they indicated similar features. The models built using the features selected by RFE combined with LASSO performed slightly better. On the other hand, the SHAP explainer showed that the features selected by RFE combined with RF had a more distributed impact on the target prediction. SHAP indicated clinical features to be more relevant than the biological ones but also showed that the combination of the two has a better impact on

the prediction of patient survival and CRC recurrence.

This study had some limitations. Initially, although several novel lncRNAs, *PCs* and miRNAs with clinical significance for CRC were found, the study was performed with TCGA data and no further experimental validation was carried out. It is also important to highlight that TCGA consists of data collected exclusively from patients in the United States. The amount of available data was a limiting factor, as only open-source data was used. CRC was treated as a single disease in our prediction, instead of dividing it into its anatomical sites (colon, rectum, and rectosigmoid junction), in order to mitigate this limiting factor. Finally, I believe that the development of this analysis with data collected from patients of other countries, such as Brazil, could give physicians a regional-specific view and better understanding of CRC-specific characteristics for each anatomical site as potentially related to the region where patients live. Research on biological and clinical features in CRC is still in development and requires further experimental studies, and a greater amount of CRC data to improve understanding.

Chapter 5

Conclusion

The elucidation of molecular mechanisms and factors that affect CRC can assist physicians in treatment and patient prognosis for the disease. This study analyzed open data from patients with CRC through bioinformatics and ML techniques to identify the biological and clinical aspects that may affect patient prognosis. First, I performed a comprehensive search to find clinical and biological information associated with patients with colon, rectum, and rectosigmoid cancer. I collected information associated with 391 colon, 85 rectum, and 69 rectosigmoid cancer patients from the TCGA database, specifically, the TGCA-COAD and TCGA-READ projects. I proposed two pipelines using the gathered patient information as input: one to identify the biological markers related to CRC prognosis, highlighting the differences between anatomical sites; and the other, to predict CRC recurrence and patient survival and to interpret the impact of clinical and biological aspects on CRC.

In the first pipeline that sought to discover CRC-related biological markers, I created a workflow with four steps: differential expression analysis; ceRNA network construction; functional analysis; and survival analysis. The result of which was the construction and analysis of ceRNA networks for colon, rectum, and rectosigmoid cancer in order to provide clinical significance and functional implications for each of these sites. Considering the functional aspects, the molecules present in the ceRNA networks suggested potential roles in known cancer pathways, such as: cell proliferation and Wnt signaling, as common mechanisms among the CRC anatomical sites. Considering the clinical aspects, I assessed the impact of these molecules on patient survival. In conclusion, this method allowed for the identification of biomarkers with a potential role in CRC prognosis, namely, hsa-miR-1271-5p, NRG1, hsa-miR-130a-3p, SNHG16, and hsa-miR-495-3p, in the colon; E2F8, in the rectum; and of DMD and hsa-miR-130b-3p, in the rectosigmoid junction.

In the second pipeline, I created a workflow with two steps: data pre-processing; and model construction to build a ML model to predict CRC recurrence and patient survival.

The use of LR and RF resulted in the best accuracy of 90% and 83% for predicting patient survival and CRC recurrence respectively. The use of the six proposed ML algorithms also showed overall good performance, specifically, RF displayed good overall results, which was also highlighted in other studies [9, 10, 11]. Furthermore, results of this thesis suggest that the combination of biological and clinical features may help to predict patient prognosis. The biological features with greatest average impact in all cases, namely, SNHG16, hsa-miR-130b-3p, hsa-miR-495-3p and KCNQ1OT1 were also pointed out by other studies [169, 213, 214, 207, 189, 190, 215, 216] as important in CRC development. Results also showed that age, ethnicity, pathological stage, chemotherapy, and lymph node count, clinical features confirmed as relevant by previous studies [217, 218, 219, 220, 221] are important even when combined with biological features.

This study may confirm that which is common knowledge: Machine learning algorithms in bioinformatics can be used for prediction, classification, and feature selection to enhance interpretation of CRC characteristics. Following the proposed pipelines, physicians can better understand the underlying mechanisms of CRC at its anatomical sites, as well as use the proposed model to help predict patient prognosis. The findings of this study are a starting point for further studies on CRC, using bioinformatic and ML techniques. Also, although this study was applied to data from patients from the USA, it can be generalized, and running these pipelines in Brazilian patient's data could lead to an improvement in CRC interpretation, especially in countries with diversity and inequality in the demographic landscape, which can affect CRC prognosis.

5.1 Contributions

In this thesis, I proposed two computational methods, using bioinformatics tools and ML techniques to deepen knowledge of the underlying mechanisms of CRC. The output from the first method, generated through a pipeline, indicated several potential prognostic markers for colon, rectum, and rectosigmoid junction cancer. I also created specific ceRNA networks for each CRC anatomical site, highlighting their potential common mechanisms. The method proposed to find the biomarkers was published by Vieira et al. [169][1]. Also, I developed a ML method that uses clinical features and the biological markers found through my initial work to predict patient survival and CRC recurrence. The model achieved good accuracy and indicated several potential clinical and biological features related to patient prognosis. Finally, I built a data repository containing proteins, miRNAs, and lncRNAs related to CRC[2].

[1]This article has been cited in Pan et al. [222], Bayrak et al. [223], Chen et al. [224] and Ding et al. [225].

[2]https://github.com/lmacielvieira/crc_pipeline/tree/main/method1/supplementary_material

5.2 Future work

The amount of available data, mainly for rectosigmoid junction cancer was a limiting factor as the number of available patients with this type of cancer was low. The fact that clinical features known to be relevant in CRC development, such as weight and height, were missing for some patients, was also a limiting factor. Another important aspect of the input data was that patient information was concentrated in one country, and may vary for other countries. Thus, standardizing the collected data from patients, as well as collecting data from other countries, could improve further analysis. Given these limitations, I intend to gather more data from different databases, including data collected in hospitals that contains the features used in our pipeline, in order to build a more robust model to predict patient prognosis. I also intend to run the pipeline with patient data from Brazil, or other countries, in order to analyze and apply the methods using data for specific populations.

References

[1] *Estimativa 2020 - Incidência de câncer no brasil, INCA.* `https://www.inca.gov.br/sites/ufu.sti.inca.local/files//media/document//estimativa-2020-incidencia-de-cancer-no-brasil.pdf`. Accessed: 2023-01-27. vi, x

[2] Zhong, M. E., C. Yanyu, and G. Zhang: *Lncrna H19 regulates PI3K–Akt signal pathway by functioning as a ceRNA and predicts poor prognosis in colorectal cancer: integrative analysis of dysregulated ncRNA-associated ceRNA network.* Springer, 19(1):148, 2019. vii, 46, 58, 59, 60

[3] Gao, R., C. Fang, J. Xu, *et al.*: *Lncrna CACS15 contributes to oxaliplatin resistance in colorectal cancer by positively regulating ABCC1 through sponging miR-145.* Archives of Biochemistry and Biophysics, 663:183–191, 2019. vii, 2, 9, 15, 32, 37

[4] Huang, Q. R. and X. B. Pan: *Prognostic lncRNAs, miRNAs, and mRNAs form a competing endogenous RNA network in colon cancer.* Frontiers, 9:712, 2019. vii

[5] Zhang, Z., S. Wang, D. Ji, W. Qian, *et al.*: *Construction of a ceRNA network reveals potential lncRNA biomarkers in rectal adenocarcinoma.* Oncology reports, 39(5):2101–2113, 2018. vii

[6] Fan, Q. and B. Liu: *Comprehensive analysis of a long noncoding RNA-associated competing endogenous RNA network in colorectal cancer.* OncoTargets and therapy, 11:2453, 2018. vii

[7] Pan, H., J. Pan, S. Song, L. Ji, *et al.*: *Identification and development of long noncoding RNA-associated regulatory network in colorectal cancer.* Journal of cellular and molecular medicine, 23(8):5200–5210, 2019. vii, 46, 58, 60

[8] Zhang, H., Z. Wang, J. Wu, *et al.*: *Long noncoding RNAs predict the survival of patients with colorectal cancer as revealed by constructing an endogenous RNA network using bioinformation analysis.* Cancer medicine, 8(3):863–873, 2019. vii, 37, 41, 46, 58, 60

[9] Gründner, J., H. U. Prokosch, M. Stürzl, *et al.*: *Predicting clinical outcomes in colorectal cancer using machine learning.* In *MIE*, pages 101–105, 2018. vii, viii, xi, xiii, 3, 40, 41, 80, 81, 84

[10] Achilonu, O. J., J. Fabian, B. Bebington, *et al.*: *Predicting colorectal cancer recurrence and patient survival using supervised machine learning approach: a south*

african population-based study. Frontiers in Public Health, 9, 2021. vii, viii, xi, xiii, 3, 40, 41, 80, 81, 84

[11] Gupta, P., S. F. Chiang, P. Sahoo, S. Mohapatra, *et al.*: *Prediction of colon cancer stages and survival period with machine learning approach.* Cancers, 11(12):2007, 2019. vii, viii, xi, xiii, 3, 40, 41, 80, 81, 84

[12] Uchida, S. and J. C. Adams: *Physiological roles of non-coding RNAs.* American Journal of Physiology-Cell Physiology, 317(1):C1–C2, 2019. 1

[13] Zhang, P., W. Wu, Q. Chen, and M. Chen: *Non-coding RNAs and their integrated networks.* Journal of integrative bioinformatics, 16(3), 2019. 1

[14] López-Jiménez, E. and E. Andrés-León: *The implications of ncRNAs in the development of human diseases.* Non-coding RNA, 7(1):17, 2021. 1

[15] Gusic, M. and H. Prokisch: *ncRNAs: new players in mitochondrial health and disease?* Frontiers in genetics, 11:95, 2020. 1

[16] Slack, F. J. and A. M. Chinnaiyan: *The role of non-coding RNAs in oncology.* Cell, 179(5):1033–1055, 2019. 1

[17] Krebs, J. E, E. S. Goldstein, and S. T. Kilpatrick: *Lewin's genes XII.* Jones & Bartlett Learning, 2017. 1, 10

[18] Yang, Q., Z. L. Cui, Q. Wang, *et al.*: *PlncRNA-1 induces apoptosis through the Her-2 pathway in prostate cancer cells.* Asian Journal of Andrology, 19(4):453, 2017. 1

[19] Huang, J., N. Zhou, K. Watabe, Z. Lu, *et al.*: *Long non-coding RNA UCA1 promotes breast tumor growth by suppression of p27 (kip1).* Cell Death & Disease, 5(1):e1008, 2014. 1, 2

[20] Bridges, M. C., A. C. Daulagala, and A. Kourtidis: *LNCcation: lncRNA localization and function.* Journal of Cell Biology, 220(2), 2021. 1, 11

[21] Jiang, M. C., J. J. Ni, W. Y. Cui, B. Y. Wang, and W. Zhuo: *Emerging roles of lncRNA in cancer and therapeutic opportunities.* American journal of cancer research, 9(7):1354, 2019. 1, 2, 11

[22] Zhang, R., L. Q. Xia, W. W. Lu, J. Zhang, and J. S. Zhu: *LncRNAs and cancer.* Oncology Letters, 12(2):1233–1239, 2016. 1, 3, 13

[23] Mercer, T., M. Dinger, and J. Mattick: *Long non-coding RNAs: insights into functions.* Nature Reviews Genetics, 10(3):155–159, 2009. 1

[24] Keniry, A., D. Oxley, P. Monnier, *et al.*: *The H19 lincRNA is a developmental reservoir of miR-675 that suppresses growth and igf1r.* Nature Cell Biology, 14(7):659–665, 2012. 1

[25] Ulitsky, I., A. Shkumatava, C. Jan, *et al.*: *Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution.* Cell, 147(7):1537–1550, 2011. 1

[26] Zhang, X., S. Weissman, and P. Newburger: *Long intergenic non-coding RNA HOTAIRM1 regulates cell cycle progression during myeloid maturation in NB4 human promyelocytic leukemia cells.* RNA Biology, 11(6):777–787, 2014. 1

[27] Li, J., M. Zhang, G. An, and Q. Ma: *LncRNA TUG1 acts as a tumor suppressor in human glioma by promoting cell apoptosis.* Experimental Biology and Medicine, page 1535370215622708, 2016. 1

[28] Prensner, J. R. and A. M. Chinnaiyan: *The emergence of lncRNAs in cancer biology.* Cancer Discovery, 1(5):391–407, 2011. 2

[29] Fachel, A., A. Tahira, S. Vilella-Arias, V. Maracaja-Coutinho, E. Gimba, G. Vignal, F. Campos, E. Reis, and S. Verjovski-Almeida: *Expression analysis and in silico characterization of intronic long noncoding RNAs in renal cell carcinoma: emerging functional associations.* Mol Cancer, 12(140):10–1186, 2013. 2

[30] Paci, P., T. Colombo, and L. Farina: *Computational analysis identifies a sponge interaction network between long non-coding RNAs and messenger RNAs in human breast cancer.* BMC Systems Biology, 8(1):83, 2014. 2, 13, 15, 30

[31] Beckedorff, F. C., M. S. Amaral, C. Deocesano-Pereira, and S. Verjovski-Almeida: *Long non-coding RNAs and their implications in cancer epigenetics.*, 2013. 2

[32] Reis, E. and S. Verjovski-Almeida: *Perspectives of long non-coding RNAs in cancer diagnostics.* Frontiers in Genetics, 3:32, 2012. 2

[33] Vorvis, C., M. Hatziapostolou, S. Mahurkar-Joshi, *et al.*: *Transcriptomic and CRISPR/Cas9 technologies reveal FOXA2 as a tumor suppressor gene in pancreatic cancer.* American Journal of Physiology-Gastrointestinal and Liver Physiology, 310(11):G1124–G1137, 2016. 2

[34] Fang, L., J. Sun, Z. Pan, *et al.*: *Long non-coding RNA NEAT1 promotes hepatocellular carcinoma cell proliferation through the regulation of miR-129-5p-VCP-ικb.* American Journal of Physiology-Gastrointestinal and Liver Physiology, 313(2):G150–G156, 2017. 2

[35] Zhang, X. F., T. Liu, Y. Li, and S. Li: *Overexpression of long non-coding RNA CCAT1 is a novel biomarker of poor prognosis in patients with breast cancer.* International Journal of Clinical and Experimental Pathology, 8(8):9440, 2015. 2

[36] Miao, Y., R. Fan, L. Chen, and H. Qian: *Clinical significance of long non-coding RNA MALAT1 expression in tissue and serum of breast cancer.* Annals of Clinical & Laboratory Science, 46(4):418–424, 2016. 2

[37] Siegel, R., J. Ma, Z. Zou, and A. Jemal: *Cancer statistics, 2014.* CA: a Cancer Journal for Clinicians, 64(1):9–29, 2014. 2

[38] Hanahan, D. and R. A. Weinberg: *Hallmarks of cancer: the next generation.* Cell, 144(5):646–674, 2011. 2, 6

[39] Ning, S., J. Zhang, P. Wang, H. Zhi, J. Wang, Y. Liu, Y. Gao, M. Guo, M. Yue, L. Wang, *et al.*: *Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers.* Nucleic Acids Research, 44(D1):D980–D985, 2015. 2, 3

[40] Ye, J., J. Li, and P. Zhao: *Roles of ncRNAs as ceRNAs in Gastric cancer.* Genes, 12(7):1036, 2021. 2, 3

[41] Xu, J., J. Xu, X. Liu, and J. Jiang: *The role of lncRNA-mediated ceRNA regulatory networks in pancreatic cancer.* Cell death discovery, 8(1):1–11, 2022. 2

[42] Aprile, M., V. Costa, A. Cimmino, and G. A. Calin: *Emerging role of oncogenic long noncoding RNA as cancer biomarkers.* International journal of cancer, 152(5):822–834, 2023. 2, 15

[43] Rebbeck, T. R.: *Prostate cancer genetics: variation by race, ethnicity, and geography.* Seminars in radiation oncology, 27(1):3–10, 2017. 2

[44] Freitas, S. C., D. Sanderson, S. Caspani, *et al.*: *New frontiers in Colorectal Cancer Treatment Combining nanotechnology with Photo-and Radiotherapy.* Cancers, 15(2):383, 2023. 2, 6

[45] Ding, J., J. Zhao, Z.and Song, *et al.*: *MiR-223 promotes the doxorubicin resistance of colorectal cancer cells via regulating epithelial–mesenchymal transition by targeting FBXW7.* Acta Biochimica et Biophysica Sinica, 50(6):597–604, 2018. 2, 35, 37

[46] Thorenoor, N., P. Faltejskova-Vychytilova, S. Hombach, *et al.*: *Long non-coding RNA ZFAS1 interacts with CDK1 and is involved in p53-dependent cell cycle control and apoptosis in colorectal cancer.* Oncotarget, 7(1):622, 2016. 2, 39, 41

[47] Wang, Q., H. Zhang, X. Shen, and S. Ju: *Serum microRNA-135a-5p as an auxiliary diagnostic biomarker for colorectal cancer.* Annals of Clinical Biochemistry, 54(1):76–85, 2017. 2, 33, 37

[48] Inoue, A., H. Yamamoto, M. Uemura, *et al.*: *MicroRNA-29b is a novel prognostic marker in colorectal cancer.* Annals of Surgical Oncology, 22(3):1410–1418, 2015. 2, 33, 37

[49] Conte, F., G. Fiscon, P. Sibilio, *et al.*: *An overview of the computational models dealing with the regulatory ceRNA mechanism and ceRNA deregulation in cancer.* Pseudogenes, pages 149–164, 2021. 2

[50] Peng, W. X., P. Koirala, and Y. Y. Mo: *LncRNA-mediated regulation of cell signaling in cancer.* Oncogene, 36(41):5661, 2017. 3, 13

[51] Ma, C., K. Nong, H. Zhu, W. Wang, X. Huang, Z. Yuan, and K. Ai: *H19 promotes pancreatic cancer metastasis by derepressing let-7's suppression on its target HMGA2-mediated EMT.* Tumor Biology, 35(9):9163–9169, 2014. 3, 13

[52] Zheng, H. T., D. B. Shi, Y. W. Wang, X. X. Li, Y. Xu, P. Tripathi, W. L. Gu, G. X. Cai, and S. J. Cai: *High expression of lncRNA MALAT1 suggests a biomarker of poor prognosis in colorectal cancer.* International Journal of Clinical and Experimental Pathology, 7(6):3174, 2014. 3, 13

[53] Chakravarty, D., A. Sboner, S. S. Nair, E. Giannopoulou, R. Li, S. Hennig, J. M. Mosquera, J. Pauwels, K. Park, M. Kossai, *et al.*: *The oestrogen receptor alpha-regulated lncRNA NEAT1 is a critical modulator of prostate cancer.* Nature Communications, 5:5383, 2014. 3

[54] Xiang, J. F., Q. F. Yin, T. Chen, Y. Zhang, X. O. Zhang, Z. Wu, S. Zhang, H. B. Wang, J. Ge, X. Lu, *et al.*: *Human colorectal cancer-specific CCAT1-L lncRNA regulates long-range chromatin interactions at the MYC locus.* Cell Research, 24(5):513, 2014. 3

[55] Chen, G., Z. Wang, D. Wang, C. Qiu, M. Liu, X. Chen, Q. Zhang, G. Yan, and Q. Cui: *Lncrnadisease: a database for long-non-coding RNA-associated diseases.* Nucleic Acids Research, 41(D1):D983–D986, 2012. 3

[56] Gong, J., W. Liu, J. Zhang, X. Miao, and A. Y. Guo: *lncRNASNP: a database of SNPs in lncRNAs and their potential functions in human and mouse.* Nucleic Acids Research, 43(D1):D181–D186, 2014. 3

[57] Ning, S., Z. Zhao, J. Ye, P. Wang, H. Zhi, R. Li, T. Wang, and X. Li: *LincSNP: a database of linking disease-associated SNPs to human large intergenic non-coding RNAs.* BMC Bioinformatics, 15(1):152, 2014. 3

[58] Griffiths-Jones, S., H. K. Saini, S. van Dongen, and A. J. Enright: *miRBase: tools for microRNA genomics.* Nucleic Acids Research, 36(suppl_1):D154–D158, 2007. 3

[59] *TCGA research network.* `https://www.cancergenome.nih.gov`. Accessed: 2015-11-15. 3, 43

[60] Rhodes, D. R, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pander, and A. M Chinnaiyan: *ONCOMINE: a cancer microarray database and integrated data-mining platform.* Neoplasia, 6(1):1–6, 2004. 3

[61] Bamford, S., E. Dawson, S. Forbes, J. Clements, *et al.*: *The cosmic (catalogue of somatic mutations in cancer) database and website.* British Journal of Cancer, 91(2):355–358, 2004. 3

[62] Xie, B., Q. Ding, H. Han, and D. Wu: *miRCancer: a microRNA-cancer association database constructed by text mining on literature.* Bioinformatics, 29(5):638–644, 2013. 3, 30, 36

[63] Gabriel, E., K. Attwood, E. Al-Sukhni, *et al.*: *Age-related rates of colorectal cancer and the factors associated with overall survival.* Journal of gastrointestinal oncology, 9(1):96, 2018. 3

[64] Yang, Y., G. Wang, J. He, *et al.*: *Gender differences in colorectal cancer survival: a meta-analysis*. International journal of cancer, 141(10):1942–1949, 2017. 3

[65] Bertram, J. S.: *The molecular biology of cancer*. Molecular Aspects of Medicine, 21(6):167–223, 2000. 5, 6

[66] DeBerardinis, R. J., J. J. Lum, G. Hatzivassiliou, and C. B. Thompson: *The biology of cancer: metabolic reprogramming fuels cell growth and proliferation*. Cell Metabolism, 7(1):11–20, 2008. 5

[67] Ferreira, L. LG and A. D. Andricopulo: *Cancer estimates in Brazil reveal progress for the most lethal malignancies*. Current Topics in Medicinal Chemistry, 20(22):1962–1966, 2020. 6

[68] Gao, Y., S. Shang, S. Guo, *et al.*: *Lnc2Cancer 3.0: an updated resource for experimentally supported lncrna/circrna cancer associations and web tools based on RNA-seq and scRNA-seq data*. Nucleic acids research, 49(D1):D1251–D1258, 2021. 6, 29

[69] Tahira, A. C., M. S. Kubrusly, M. F. Faria, B. Dazzani, R. S. Fonseca, V. Maracaja-Coutinho, S. Verjovski-Almeida, M. CC Machado, and E. M. Reis: *Long noncoding intronic RNAs are differentially expressed in primary and metastatic pancreatic cancer*. Molecular Cancer, 10(1):141, 2011. 6, 8, 13

[70] Christian, Wittekind, Brierley James, Lee Anne, and Eycken Elisabeth: *TNM supplement: a commentary on uniform use*. John Wiley & Sons, Hoboken, New Jersey, 2019. 6

[71] *Tractus intestinalis rectum*. `https://www.commons.wikimedia.org/wiki/File:Tractus_intestinalis_rectum.svg`. Accessed: 2021-10-27. 7

[72] Moawad, E. Y: *Clinical and pathological staging of the cancer at the nanoscale*. Cancer nanotechnology, 3(1):37–46, 2012. 9

[73] Mukai, M., K. Kishima, Ma. Yamazaki, *et al.*: *Stage ii/iii cancer of the rectosigmoid junction: an independent tumor type?* Oncology reports, 26(3):737–741, 2011. 9

[74] Zhou, J., J. Lin, H. Zhang, *et al.*: *Lncrna HAND2-AS1 sponging miR-1275 suppresses colorectal cancer progression by upregulating KLF14*. Biochemical and Biophysical Research Communications, 503(3):1848–1853, 2018. 9, 31, 37

[75] Crick, F.: *Francis crick*. The double helix, 1951:2–3, 1953. 10

[76] Machado-Lima, A., H. Del Portillo, and A. Durham: *Computational methods in noncoding RNA research*. Journal of Mathematical Biology, 56(1-2):15–49, 2008. 10

[77] Clote, P. and R. Backofen: *Computational molecular biology: an introduction a self contained approach to bioinformatics*. Chichester Wiley, 2000. 11

[78] Di, C., Q. Zhang, Y. Chen, *et al.*: *Function, clinical application, and strategies of pre-mrna splicing in cancer.* Cell Death & Differentiation, 26(7):1181–1194, 2019. 11

[79] Kalvari, I., E. P. Nawrocki, N. Ontiveros-Palacios, *et al.*: *Rfam 14: expanded coverage of metagenomic, viral and microRNA families.* Nucleic Acids Research, 49(D1):D192–D200, 2021. 11, 12, 29

[80] Oliveira, J. V. A.: *Identificação de snoRNAs usando aprendizagem de máquina.* Master's thesis, Universidade de Brasília, 2016. 11, 12, 29

[81] Nawrocki, E. P. and S. R. Eddy: *Infernal 1.1: 100-fold faster RNA homology searches.* Bioinformatics, 29(22):2933–2935, 2013. 12

[82] Chalbatani, G. M., H. Dana, E. Gharagouzloo, S. Grijalvo, and others.: *Small interfering RNAs (siRNAs) in cancer therapy: a nano-based approach.* International journal of nanomedicine, 14:3111, 2019. 12

[83] Yao, Q., Y. Chen, and X. Zhou: *The roles of microRNAs in epigenetic regulation.* Current opinion in chemical biology, 51:11–17, 2019. 12

[84] Pedroza-Torres, A., S. L Romero-Córdoba, M. Justo-Garrido, *et al.*: *MicroRNAs in tumor cell metabolism: roles and therapeutic opportunities.* Frontiers in Oncology, 9:1404, 2019. 12, 13

[85] *Created with BioRender.com.* `https://www.biorender.com/`. Accessed: 2022-12-27. 12

[86] Cui, M., H. Wang, Xi. Yao, and others.: *Circulating microRNAs in cancer: potential and challenge.* Frontiers in genetics, 10:626, 2019. 13

[87] Dahariya, S., I. Paddibhatla, S. Kumar, and other.: *Long non-coding RNA: Classification, biogenesis and functions in blood cells.* Molecular immunology, 112:82–92, 2019. 13, 15

[88] Vieira, L., C. Grativol, F. Thiebaut, T. Carvalho, P. Hardoim, A. Hemerly, S. Lifschitz, P. C. Ferreira, and M. E. MT Walter: *PlantRNA_sniffer: A SVM-based workflow to predict long intergenic non-coding RNAs in plants.* Non-Coding RNA, 3(1):11, 2017. 14, 19

[89] Yousefi, H., M. Maheronnaghsh, F. Molaei, *et al.*: *Long noncoding RNAs and exosomal lncRNAs: classification, and mechanisms in breast cancer metastasis and drug resistance.* Oncogene, 39(5):953–974, 2020. 13, 15

[90] Cesana, M., D. Cacchiarelli, I. Legnini, *et al.*: *A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA.* Cell, 147(2):358–369, 2011. 15, 16

[91] Poliseno, L., L. Salmena, J. Zhang, B. Carver, W. J Haveman, and P. P. Pandolfi: *A coding-independent function of gene and pseudogene mRNAs regulates tumour biology.* Nature, 465(7301):1033, 2010. 15

[92] Sumazin, P., X. Yang, H. S. Chiu, *et al.*: *An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma.* Cell, 147(2):370–381, 2011. 15

[93] Chan, J. and Y. Tay: *Noncoding RNA: RNA regulatory networks in cancer.* International journal of molecular sciences, 19(5):1310, 2018. 16

[94] Zhang, J. J., X. H. Zhou, Y. Zhou, *et al.*: *Bufalin suppresses the migration and invasion of prostate cancer cells through HOTAIR, the sponge of miR-520b.* Acta Pharmacologica Sinica, page 1, 2019. 15

[95] Wang, W., W. Lou, B. Ding, *et al.*: *A novel mRNA-miRNA-lncRNA competing endogenous RNA triple sub-network associated with prognosis of pancreatic cancer.* Aging (Albany NY), 11(9):2610, 2019. 15

[96] Zhang, F., C. Wu, J. Zhang, Z. Huang, *et al.*: *Identification of ceRNA-based H19/SIX4 regulatory axis as a prognostic biomarker for colorectal cancer via high throughput transcriptomic data.* Preprint at https://doi.org/10.21203/rs.3.rs-2233353/v1, 2022. 15

[97] Lin, X., S. Zhuang, X. Chen, J. Du, *et al.*: *lncRNA ITGB8-AS1 functions as a ceRNA to promote colorectal cancer growth and migration through integrin-mediated focal adhesion signaling.* Molecular Therapy, 30(2):688–702, 2022. 15

[98] Li, T., W. Liu, C. Wang, M. Wang, *et al.*: *Multidimension analysis of the prognostic value, immune regulatory function, and ceRNA network of LY6e in individuals with colorectal cancer.* Journal of immunology research, 2022, 2022. 15

[99] Russell, S. and P. Norvig: *Artificial Intelligence: a modern approach*, volume 3. Pearson, 2010. 15, 17

[100] Cortes, C. and V. Vapnik: *Support-vector networks.* Machine Learning, 20(3):273–297, 1995. 16

[101] Breiman, L., JH Friedman, RA Olshen, and CJ Stone: *Classification and regression trees (1984). Monterey, ca: Wadsworth Brooks.* 17, 18

[102] Altman, N. S.: *An introduction to kernel and nearest-neighbor nonparametric regression.* The American Statistician, 46(3):175–185, 1992. 17

[103] Kaufman, L and PJ Rousseeuw: *Clustering by means of medoids [w:] statistical data analysis based on the ll-norm and related methods, red. y. dodge*, 1987. 17

[104] MacQueen, J. *et al.*: *Some methods for classification and analysis of multivariate observations.* In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA., 1967. 17

[105] Sutton, R. and A. Barto: *Reinforcement learning: an introduction. The MIT Press.* Cambridge, MA, 1998. 17

[106] Rummery, G. A. and M. Niranjan: *On-line Q-learning using connectionist systems*, volume 37. University of Cambridge, Department of Engineering, 1994. 17

[107] Boyan, J. A: *Technical update: Least-squares temporal difference learning.* Machine Learning, 49(2):233–246, 2002. 17

[108] Xiaojin, Z. and G. Zoubin: *Learning from labeled and unlabeled data with label propagation.* Technical Report CMU-CALD-02–107, Carnegie Mellon University, 2002. 17

[109] Zhou, D., O. Bousquet, T. N Lal, J. Weston, and B. Schölkopf: *Learning with local and global consistency.* In *Advances in Neural Information Processing Systems*, pages 321–328, 2004. 17

[110] Bisong, E. and E. Bisong: *Logistic regression.* Building machine learning and deep learning models on google cloud platform: A comprehensive guide for beginners, pages 243–250, 2019. 17

[111] *Logistic regression in machine learning.* `https://www.javatpoint.com/logistic-regression-in-machine-learning`. Accessed: 2023-01-28. 18

[112] Wu, X., V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J McLachlan, A. Ng, B. Liu, S. Y. Philip, *et al.*: *Top 10 algorithms in data mining.* Knowledge and Information Systems, 14(1):1–37, 2008. 18

[113] *KNN classifier approach.* `https://www.researchgate.net/figure/297728234_fig3_Figure-7k-NN-classifier-approach`. Accessed: 2023-01-22. 19

[114] Bansal, M., A. Goyal, and A. Choudhary: *A comparative analysis of K-Nearest Neighbour, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning.* Decision Analytics Journal, page 100071, 2022. 19

[115] *Interpreting Ctree output in R.* `https://stats.stackexchange.com/questions/171301/interpreting-ctree-partykit-output-in-r`. Accessed: 2023-01-22. 20

[116] Zhang, C. and Y. Ma: *Ensemble machine learning: methods and applications.* Springer, 2012. 20

[117] Zhou, Z. H.: *Ensemble Methods: foundations and algorithms*, volume 1. CRC press, 2012. 20, 21, 26

[118] Speiser, J. Lynn, M. E Miller, J. Tooze, *et al.*: *A comparison of random forest variable selection methods for classification prediction modeling.* Expert systems with applications, 134:93–101, 2019. 21

[119] Wang, F., Z. Li, F. He, *et al.*: *Feature learning viewpoint of AdaBoost and a new algorithm.* IEEE Access, 7:149890–149899, 2019. 22

[120] Storcheus, D., A. Rostamizadeh, and S. Kumar: *A survey of modern questions and challenges in feature extraction.* In *Feature Extraction: Modern Questions and Challenges*, pages 1–18. PMLR, 2015. 23, 80

[121] Bolón-Canedo, V. and A. Alonso-Betanzos: *Ensembles for feature selection: A review and future trends.* Information Fusion, 52:1–12, 2019. 24

[122] Jeon, H. and S. Oh: *Hybrid-recursive feature elimination for efficient feature selection.* Applied Sciences, 10(9):3211, 2020. 24

[123] Ranstam, J. and J.A Cook: *Lasso regression.* Journal of British Surgery, 105(10):1348–1348, 2018. 24

[124] Slack, D., S. Hilgard, E. Jia, *et al.*: *Fooling lime and shap: Adversarial attacks on post hoc explanation methods.* In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020. 24

[125] Chawla, N. V., K. W. Bowyer, O. Hall, and W. P. Kegelmeyer: *SMOTE: synthetic minority over-sampling technique.* Journal of Artificial Intelligence Research, 16:321–357, 2002. 25

[126] Rodrigues, T.: *The good, the bad, and the ugly in chemical and biological data for machine learning.* Drug Discovery Today: Technologies, 32:3–8, 2019. 25, 26

[127] Xu, C. and S. A Jackson: *Machine learning and complex biological data*, 2019. 26

[128] Ghojogh, B. and M. Crowley: *The theory behind overfitting, cross validation, regularization, bagging, and boosting: tutorial.* arXiv preprint arXiv:1905.12787, 2019. 27

[129] Liashchynskyi, P. and P. Liashchynskyi: *Grid search, random search, genetic algorithm: a big comparison for nas.* arXiv preprint arXiv:1912.06059, 2019. 28

[130] Bernal, A., U. Ear, and N. Kyrpides: *Genomes OnLine Database (GOLD): a monitor of genome projects world-wide.* Nucleic Acids Research, 29(1):126–127, 2001. `http://www.genomesonline.org/cgi-bin/GOLD/index.cgi`. 29

[131] Sabin, L., M. Delás, and G. Hannon: *Dogma derailed: The many influences of RNA on the genome.* Molecular Cell, 49(5):783–794, 2013. 29

[132] Xiong, J.: *Essential Bioinformatics.* Cambridge University Press, 2006. 29

[133] Zhao, L., J. Wang, Y. Li, *et al.*: *NONCODEV6: an updated database dedicated to long non-coding rna annotation in both animals and plants.* Nucleic acids research, 49(D1):D165–D171, 2021. 29

[134] Quek, X. C., D. W Thomson, J. LV Maag, *et al.*: *lncRNAdb v2. 0: expanding the reference database for functional long noncoding RNAs.* Nucleic acids research, 43(D1):D168–D173, 2015. 29

[135] Volders, P. J., J. Anckaert, K. Verheggen, *et al.*: *LNCipedia 5: towards a reference set of human long non-coding RNAs.* Nucleic acids research, 47(D1):D135–D139, 2019. 29

[136] Kozomara, A., M. Birgaoanu, and S. Griffiths-Jones: *mirbase: from microRNA sequences to function.* Nucleic acids research, 47(D1):D155–D162, 2019. 29

[137] Bao, Z., Z. Yang, Z. Huang, *et al.*: *Lncrnadisease 2.0: an updated database of long non-coding rna-associated diseases.* Nucleic acids research, 47(D1):D1034–D1037, 2019. 29

[138] Barrett, T., T. Suzek, D. Troup, *et al.*: *NCBI GEO: mining millions of expression profiles — database and tools.* Nucleic Acids Research, 33(suppl_1):D562–D566, 2005. 30

[139] Katarzyna, Tomczak, Czerwińska Patrycja, and Wiznerowicz Maciej: *The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge.* Contemporary oncology, 19(1A):A68, 2015. 30

[140] Franco-Zorrilla, J., A. Valli, M. Todesco, *et al.*: *Target mimicry provides a new mechanism for regulation of microRNA activity.* Nature Genetics, 39(8):1033, 2007. 30

[141] Agarwal, V., G. W Bell, J. W. Nam, and D. P Bartel: *Predicting effective microRNA target sites in mammalian mRNAs.* Elife, 4:e05005, 2015. 30

[142] Betel, D., A. Koppal, P. Agius, C. Sander, and C. Leslie: *Comprehensive modeling of microrna targets predicts functional non-conserved and non-canonical sites.* Genome Biology, 11(8):R90, 2010. 30

[143] Li, J. H., S. Liu, H. Zhou, L. H. Qu, and J. H. Yang: *starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data.* Nucleic Acids Research, 42(D1):D92–D97, 2013. 30

[144] Han, Y., Y. N. Yang, H. H. Yuan, *et al.*: *UCA1, a long non-coding RNA up-regulated in colorectal cancer influences cell proliferation, apoptosis and cell cycle distribution.* Pathology, 46(5):396–401, 2014. 31, 37

[145] Zhong, F., W. Zhang, and Y.and others Cao: *Lncrna NEAT1 promotes colorectal cancer cell proliferation and migration via regulating glial cell-derived neurotrophic factor by sponging miR-196a-5p.* Acta Biochimica et Biophysica Sinica, 50(12):1190–1199, 2018. 31, 37

[146] Lu, X., Y. Yu, and S. Tan: *Long non-coding XIAP-AS1 regulates cell proliferation, invasion and cell cycle in colon cancer.* Artificial Cells, Nanomedicine, and Biotechnology, 47(1):767–775, 2019. 32, 37

[147] Ke, T. W., P. L. Wei, K. T. Yeh, *et al.*: *MiR-92a promotes cell metastasis of colorectal cancer through PTEN-mediated PI3K/AKT pathway.* Annals of Surgical Oncology, 22(8):2649–2655, 2015. 33, 37

[148] Igarashi, H., H. Kurihara, K. Mitsuhashi, *et al.*: *Association of microRNA-31-5p with clinical efficacy of anti-EGFR therapy in patients with metastatic colorectal cancer.* Annals of Surgical Oncology, 22(8):2640–2648, 2015. 33, 37

[149] Zu, C., T. Liu, and G. Zhang: *MicroRNA-506 inhibits malignancy of colorectal carcinoma cells by targeting LAMC1.* Annals of Clinical & Laboratory Science, 46(6):666–674, 2016. 34, 37

[150] Ozawa, T., T. Matsuyama, Y. Toiyama, *et al.*: *CCAT1 and CCAT2 long noncoding RNAs, located within the 8q. 24.21 gene desert, serve as important prognostic biomarkers in colorectal cancer.* Annals of Oncology, 28(8):1882–1888, 2017. 34, 37

[151] Dou, J., Y. Ni, X. He, *et al.*: *Decreasing lncRNA HOTAIR expression inhibits human colorectal cancer stem cells.* American Journal of Translational Research, 8(1):98, 2016. 34, 37

[152] Ma, S., D. Yang, Y. Liu, *et al.*: *LncRNA BANCR promotes tumorigenesis and enhances adriamycin resistance in colorectal cancer.* Aging (Albany NY), 10(8):2062, 2018. 35, 37

[153] Lee, H., C. Kim, J. L. Ku, W. Kim, *et al.*: *A long non-coding RNA snar contributes to 5-fluorouracil resistance in human colon cancer cells.* Molecules and Cells, 37(7):540, 2014. 35, 37

[154] Yin, D., X. He, E. Zhang, *et al.*: *Long noncoding rna GAS5 affects cell proliferation and predicts a poor prognosis in patients with colorectal cancer.* Medical Oncology, 31(11):253, 2014. 35, 37

[155] Han, X., Li. Wang, and Y.and others Ning: *Long non-coding RNA AFAP1-AS1 facilitates tumor growth and promotes metastasis in colorectal cancer.* Biological Research, 49(1):36, 2016. 36, 37

[156] Yuan, W., X. Li, L. Liu, *et al.*: *Comprehensive analysis of lncRNA-associated ceRNA network in colorectal cancer.* Biochemical and Biophysical Research Communications, 508(2):374–379, 2019. 36, 41

[157] Robinson, M., D. McCarthy, and G. Smyth: *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 26(1):139–140, 2010. 36

[158] Jeggari, A., D. Marks, and E. Larsson: *mircode: a map of putative microRNA target sites in the long non-coding transcriptome.* Bioinformatics, 28(15):2062–2063, 2012. 36

[159] Huang, D., B. Sherman, and R. Lempicki: *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nature Protocols, 4(1):44, 2009. 36

[160] Lin, H. and D. Zelterman: *Modeling survival data: extending the Cox model*, 2002. 36

[161] Yu, G., L. G. Wang, Y. Han, and Q. Y. He: *clusterProfiler: an R package for comparing biological themes among gene clusters.* Omics: A Journal of Integrative Biology, 16(5):284–287, 2012. 37

[162] Subramanian, A., P. Tamayo, V. Mootha, *et al.*: *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proceedings of the National Academy of Sciences, 102(43):15545–15550, 2005. 37

[163] Falzone, L., L. Scola, A. Zanghì, A. Biondi, *et al.*: *Integrated analysis of colorectal cancer microRNA datasets: Identification of microRNAs associated with tumor development.* Aging (Albany NY), 10(5):1000, 2018. 38, 41

[164] Bhome, R., R. W Goh, M. Bullock, *et al.*: *Exosomal microRNAs derived from colorectal cancer-associated fibroblasts: role in driving cancer progression.* Aging (Albany NY), 9(12):2666, 2017. 38, 41

[165] Hu, Y., H. Y. Chen, C. Y. Yu, *et al.*: *A long non-coding RNA signature to improve prognosis prediction of colorectal cancer.* Oncotarget, 5(8):2230, 2014. 39, 41

[166] Gentleman, R., V. Carey, D. Bates, *et al.*: *Bioconductor: open software development for computational biology and bioinformatics.* Genome Biology, 5(10):R80, 2004. 39

[167] Qiu, J. j. and J. b. Yan: *Long non-coding rna LINC01296 is a potential prognostic biomarker in patients with colorectal cancer.* Tumor Biology, 36(9):7175–7183, 2015. 39, 41

[168] Yang, J., Ji. Lin, T. Liu, *et al.*: *Analysis of lncRNA expression profiles in non-small cell lung cancers (NSCLC) and their clinical subtypes.* Lung Cancer, 85(2):110–115, 2014. 39

[169] Vieira, L. M., N. A. N. Jorge, J. B. de Sousa, J. C. Setubal, P. F. Stadler, and M. E. M. T. Walter: *Competing endogenous rna in colorectal cancer: an analysis for colon, rectum, and rectosigmoid junction.* Frontiers in oncology, page 1670, 2021. 42, 50, 51, 55, 57, 62, 64, 81, 84

[170] Li, R., H. Qu, S. Wang, and J. Wei: *GDCRNATools: an R/Bioconductor package for integrative analysis of lncRNA, miRNA and mRNA data in GDC.* Bioinformatics, 34(14):2515–2517, 2018. `https://www.academic.oup.com/bioinformatics/article-abstract/34/14/2515/4917355`. 43, 63, 68

[171] Ritchie, M., B. Phipson, D. Wu, and others.: *Limma powers differential expression analyses for rna-sequencing and microarray studies.* Nucleic acids research, 43(7):e47–e47, 2015. `https://www.academic.oup.com/nar/article/43/7/e47/2414268`. 43, 63

[172] Furio, P., S. Tarazona, T. Gabaldon, *et al.*: *spongeScan: A web for detecting microRNA binding elements in lncRNA sequences.* Nucleic acids research, 44(W1):W176–W180, 2016. 43

[173] Li, J. H., S. Liu, H. Zhou, L. H. Qu, and J. H. Yang: *starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data.* Nucleic acids research, 42(D1):D92–D97, 2014. 44

[174] Ashburner, M., C. Ball, J. Blake, D. Botstein, and others.: *Gene ontology: tool for the unification of biology.* Nature genetics, 25(1):25–29, 2000. 44

[175] Kanehisa, M. and S.. Goto: *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic acids research, 28(1):27–30, 2000. 44

[176] Schriml, L., E. Mitraka, J. Munro, B. Tauber, and others.: *Human Disease Ontology 2018 update: classification, content and workflow expansion.* Nucleic acids research, 47(D1):D955–D962, 2019. 44

[177] Zhu, Y., B. Li, Z. Liu, L. Jiang, G. Wang, *et al.*: *Long noncoding MAGI2-AS3 promotes colorectal cancer progression through regulating mir-3163/tmem106b axis.* Journal of cellular physiology, 235(5):4824–4833, 2019. 46, 58

[178] Li, M., Z. Bian, G. Jin, J. Zhang, and others.: *Lnc RNA-SNHG 15 enhances cell proliferation in colorectal cancer by inhibiting mir-338-3p.* Cancer medicine, 8(5):2404–2413, 2019. 47, 58

[179] Li, M., Z. Bian, G. Jin, and others.: *Long noncoding RNA SNHG15 enhances the development of colorectal carcinoma via functioning as a ceRNA through mir-141/sirt1/wnt/β-catenin axis.* Artificial cells, nanomedicine, and biotechnology, 47(1):2536–2544, 2019. 47, 58

[180] X, M., X. Chen, K. Lin, K. Zeng, and others.: *The long noncoding RNA SNHG1 regulates colorectal cancer cell growth through interactions with EZH2 and mir-154-5p.* Molecular cancer, 17(1):1–16, 2019. 47, 58

[181] Zhu, Y., B. Li, Z. Liu, L. Jiang, G. Wang, *et al.*: *Up-regulation of lncrna snhg1 indicates poor prognosis and promotes cell proliferation and metastasis of colorectal cancer by activation of the wnt/β-catenin signaling pathway.* Oncotarget, 8(67):111715, 2017. 47, 58

[182] Zhao, B., Z. Wan, X. Zhang, and Y. Zhao: *Comprehensive analysis reveals a four-gene signature in colorectal cancer.* Translational Cancer Research, 9(3):1395, 2020. 47

[183] Sun, Z., C. Liu, and S. Cheng: *Identification of four novel prognosis biomarkers and potential therapeutic drugs for human colorectal cancer by bioinformatics analysis.* Journal of Biomedical Research, 35(1):21, 2021. 47

[184] Wu, X., J. Cai, Z. Zuo, and J. Li: *Collagen facilitates the colorectal cancer stemness and metastasis through an integrin/pi3k/akt/snail signaling pathway.* Biomedicine & Pharmacotherapy, 114:108708, 2019. 53

[185] Le, C.C, A. Bennasroune, B. Langlois, *et al.*: *Functional interplay between collagen network and cell behavior within tumor microenvironment in colorectal cancer.* Frontiers in Oncology, 10:527, 2020. 53

[186] Sun, W., W. Nie, Z. Wang, *et al.*: *Lnc HAGLR promotes colon cancer progression through sponging mir-185-5p and activating CDK4 and CDK6 in vitro and in vivo.* OncoTargets and therapy, 13(52):5913, 2020. 58

[187] Tuffery-Giraud, S., J. Miro, M. Koenig, and M. Claustres: *Normal and altered pre-mrna processing in the DMD gene.* Human genetics, 136(9):1155–1172, 2017. 59

[188] Chae, Y. C., J. Y. Kim, J. W. Park, *et al.*: *Foxo1 degradation via g9a-mediated methylation promotes cell proliferation in colon cancer.* Nucleic acids research, 47(4):1692–1705, 2019. 59

[189] Colangelo, T., A. Fucci, C. Votino, *et al.*: *MicroRNA-130b promotes tumor development and is associated with poor prognosis in colorectal cancer.* Neoplasia, 15(9):1086–1099, 2013. 59, 81, 84

[190] Zhao, Y., G. Miao, Y. Li, *et al.*: *Microrna 130b suppresses migration and invasion of colorectal cancer cells through downregulation of integrin $\beta 1$.* PLoS One, 9(2):e87938, 2014. 59, 81, 84

[191] Zhang, Z., J. Li, Y. Huang, *et al.*: *Upregulated mir-1258 regulates cell cycle and inhibits cell proliferation by directly targeting E2F8 in CRC.* Cell proliferation, 51(6):e12505, 2018. 59, 60

[192] Lin, Y. C., Y. Wang, R. Hsu, and others.: *PCNA-mediated stabilization of E3 ligase RFWD3 at the replication fork is essential for dna replication.* Proceedings of the National Academy of Sciences, 115(52):13282–13287, 2018. 59

[193] Yan, P. y. and X. a. Zhang: *Knockdown of E2F8 suppresses cell proliferation in colon cancer cells by modulating the nf-κb pathway.* Annals of Clinical & Laboratory Science, 49(4):474–480, 2019. 60, 64

[194] Shu, P., J. Wu, Y. Tong, *et al.*: *Gene pair based prognostic signature for colorectal colon cancer.* Medicine, 97(42):2–3, 2018. 60

[195] Xu, J., J. Z, and R. Z: *Four microRNAs signature for survival prognosis in colon cancer using TCGA data.* Scientific reports, 6(38306):6–7, 2016. 60

[196] Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, and B. andothers Thirion: *Scikit-learn: Machine learning in Python.* Journal of Machine Learning Research, 12:2825–2830, 2011. 61

[197] Shimizu, D., T. M, K. Sato, and others.: *CRMP5-associated GTpase (CRAG) is a candidate driver gene for colorectal cancer carcinogenesis.* Anticancer research, 39(1):99–106, 2019. 64

[198] Bai, R., D. Wu, Z. Shi, W. Hu, *et al.*: *Pan-cancer analyses demonstrate that ANKRD6 is associated with a poor prognosis and correlates with M2 macrophage infiltration in colon cancer.* Chinese Journal of Cancer Research, 33(1):93, 2021. 64

[199] Liu, C., W. Wu, W. Chang, W., *et al.*: *miR-31-5p-DMD axis as a novel biomarker for predicting the development and prognosis of sporadic early-onset colorectal cancer.* Oncology Letters, 23(5):1–13, 2022. 64

[200] Wang, D., L. Yang, W. Yu, Q. Wu, *et al.*: *Colorectal cancer cell-derived CCL20 recruits regulatory t cells to promote chemoresistance via FOXO1/CEBPB/nf-κb signaling.* Journal for immunotherapy of cancer, 7(1):1–15, 2019. 64

[201] Jonna, S., R. A. Feldman, J. Swensen, *et al.*: *Detection of NRG1 gene fusions in solid tumors NRG1 fusions in solid tumors.* Clinical Cancer Research, 25(16):4966–4972, 2019. 64

[202] Du, F., J. Chen, H. Liu, Y. Cai, *et al.*: *SOX12 promotes colorectal cancer cell proliferation and metastasis by regulating asparagine synthesis.* Cell death & disease, 10(3):239, 2019. 64

[203] Luan, C., Y. Li, Z. Liu, and C. Zhao: *Long noncoding RNA MALAT1 promotes the development of colon cancer by regulating miR-101-3p/STC1 axis.* OncoTargets and therapy, 13:3653, 2020. 64

[204] Zhang, X. W., S. L. Li, D. Zhang, *et al.*: *RP11-619L19. 2 promotes colon cancer development by regulating the mir-1271-5p/cd164 axis.* Oncology reports, 44(6):2419–2428, 2020. 64

[205] Song, G. L., M. Xiao, X. Y. Wan, J. D., *et al.*: *Mir-130a-3p suppresses colorectal cancer growth by targeting wnt family member 1 (wnt1).* Bioengineered, 12(1):8407–8418, 2021. 64

[206] Song, D., Q. Zhang, H. Zhang, L. Zhan, and X. Sun: *Mir-130b-3p promotes colorectal cancer progression by targeting CHD9.* Cell Cycle, 21(6):585–601, 2022. 64

[207] Zhang, J. L., H. F. Zheng, K. Li, and Y. P. Zhu: *mir-495-3p depresses cell proliferation and migration by downregulating hmgb1 in colorectal cancer.* World journal of surgical oncology, 20(1):1–14, 2022. 64, 81, 84

[208] Zheng, Z. H., H. Y. You, Y. J. Feng, *et al.*: *LncRNA KCNQ1OT1 is a key factor in the reversal effect of curcumin on cisplatin resistance in the colorectal cancer cells.* Molecular and Cellular Biochemestry, 476(7):2575–2585, 2021. 64

[209] Ghafouri-Fard, S., M. Esmaeili, and M. Taheri: *H19 lncrna: roles in tumorigenesis.* Biomedicine & Pharmacotherapy, 123:109774, 2020. 64

[210] Foroumadi, R., S. Rashedi, S. Asgarian, *et al.*: *Circular RNA MYLK as a prognostic biomarker in patients with cancers: A systematic review and meta-analysis.* Cancer Reports, 5(9):e1653, 2022. 64

[211] Ke, D., Q. Wang, S. Ke, L. Zou, and Q. Wang: *Long-non coding RNA SNHG16 supports colon cancer cell growth by modulating mir-302a-3p/akt axis.* Pathology & Oncology Research, 26:1605–1613, 2020. 64

[212] Cao, W., B. Zhang, and Y. Liu: *Expression of long nonencoding ribonucleic acid SNHG20 in colon cancer tissue in its influences on chemotherapeutic sensitivity of colon cancer cells.* BioMed Research International, 2022, 2022. 64

[213] Chen, Z. Y., X. Y. Wang, Y. M. Yang, *et al.*: *LncRNA SNHG16 promotes colorectal cancer cell proliferation, migration, and epithelial–mesenchymal transition through miR-124-3p/MCP-1.* Gene therapy, 29(3):193–205, 2022. 81, 84

[214] He, X., J. Ma, M. Zhang, *et al.*: *Long non-coding rna SNHG16 activates USP22 expression to promote colorectal cancer progression by sponging mir-132-3p.* OncoTargets and therapy, 13:4283, 2020. 81, 84

[215] Duan, Q., L. Cai, K. Zheng, and others.: *lncrna KCNQ1OT1 knockdown inhibits colorectal cancer cell proliferation, migration and invasiveness via thepi3k/akt pathway.* Oncology letters, 20(1):601–610, 2020. 81, 84

[216] Mini, E., A. Lapucci, G. Perrone, *et al.*: *Rna sequencing reveals pnn and kcnq1ot1 as predictive biomarkers of clinical outcome in stage iii colorectal cancer patients treated with adjuvant chemotherapy.* International journal of cancer, 145(9):2580–2593, 2019. 81, 84

[217] Sinicrope, F. A.: *Increasing incidence of early-onset colorectal cancer.* New England Journal of Medicine, 386(16):1547–1558, 2022. 81, 84

[218] Ulanja, M. B., C. Ntafam, B. D. Beutler, *et al.*: *Race, age, and sex differences on the influence of obesity on colorectal cancer sidedness and mortality: A national cross-sectional study.* Journal of Surgical Oncology, 127(1):109–118, 2023. 81, 84

[219] Kim, H. Jin and G. S. Choi: *Clinical implications of lymph node metastasis in colorectal cancer: current status and future perspectives.* Annals of Coloproctology, 35(3):109, 2019. 81, 84

[220] Grass, F., K. T. Behm, E. Duchalais, *et al.*: *Impact of delay to surgery on survival in stage i-iii colon cancer.* European Journal of Surgical Oncology, 46(3):455–461, 2020. 81, 84

[221] Chan, G. HJ and Cheng. E Chee: *Making sense of adjuvant chemotherapy in colorectal cancer.* Journal of gastrointestinal oncology, 10(6):1183, 2019. 81, 84

[222] Pan, Y., J. Li, S. Lou, *et al.*: *Down-regulated mir-130a/b attenuates rhabdomyosarcoma proliferation via pparg.* Frontiers in molecular biosciences, 8:766887, 2022. 84

[223] Bayrak, T., Z. Çetin, E. I. Saygılı, and H. Ogul: *Identifying the tumor location-associated candidate genes in development of new drugs for colorectal cancer using machine-learning-based approach.* Medical & Biological Engineering & Computing, pages 1–21, 2022. 84

[224] Chen, X., H. Lei, Y. Cheng, *et al.*: *CXCL8, MMP12, and MMP13 are common biomarkers of periodontitis and oral squamous cell carcinoma.* Oral Diseases, 2022. 84

[225] Ding, X., L. Chen, D. Xu, *et al.*: *Pan-cancer analysis of bub1b/hsa-mir-130a-3p axis and identification of circulating hsa-mir-130a-3p as a potential biomarker for cancer risk assessment.* Evidence-based Complementary and Alternative Medicine: eCAM, 2022, 2022. 84

Annex I

Software and Data Availability

Method 1: Competing endogenous RNAs in CRC

`https://github.com/lmacielvieira/crc_pipeline/tree/main/method1`

Method 2: A biological and clinical feature analysis to predict recurrence and patient survival in CRC

`https://github.com/lmacielvieira/crc_pipeline/tree/main/method2`

Data Availability

The primary data derived from the model analysis are available for review, and replicability. The results shown here are in whole or part based upon data generated by the TCGA Research Network: `https://www.cancer.gov/tcga`, so if you use it, make sure to also cite TCGA. The data is available at: `https://github.com/lmacielvieira/crc_pipeline` or at following QR code: