FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

A Machine Learning Approach for Predicting Microsatellite Instability using RNAseq

José Miguel da Costa Simões

U. PORTO

FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

Mestrado em Engenharia Informática e Computação

Supervisor: Tânia Pereira, PhD

Co-Supervisor: Helder Oliveira, PhD

Co-Supervisor: Prof. José Carlos Machado, PhD

October 29, 2022

A Machine Learning Approach for Predicting Microsatellite Instability using RNAseq

José Miguel da Costa Simões

Mestrado em Engenharia Informática e Computação

October 29, 2022

Abstract

Cancer is a leading cause of death worldwide, having provoked more than 19 million new diagnostics and almost 10 million deaths in 2020. The most common methods to battle the disease are radio and chemotherapy. However, they leave a significant mark on patients, with side effects such as hair loss, appetite change, fatigue, or diarrhoea. Immunotherapy is a new method that is revolutionising cancer treatment. Besides being in an early stage of development, immunotherapy is one of the most promising methods to treat cancer while allowing the patients to maintain a better quality of life during the treatment and have a higher life expectancy.

The addressed problem is the prediction of Microsatellite Instability, an essential biomarker with significant prospects for his capacity to envision the immune system response. This will lead to the search for competent genes in cancer patient's immune system with the capacity to fight the disease. To find an answer to this problem, RNAseq data will be used to extract mutation and gene expression signatures, which will allow the stratification of cancer patients to define better treatment plans.

Using TCGA data, three different approaches were developed to predict MSI with several feature selection methods tested in a Multi-Layer Perceptron, a Random Forest and a K-Nearest Neighbours. The main goal was to understand the capacity of the different models to predict MSI and how each method selected the most relevant genes to make the prediction. The study aimed to confirm the capacity of RNAseq to predict MSI and to compare the use of DL models with other ML models.

The study concluded that the Multi-Layer Perceptron has a better capacity to use RNAseq data to predict MSI, with the approach that merged patients with low instability and stability on their microsatellites in the colon adenocarcinoma obtaining a performance of 98.44% of AUC and 92.67% of accuracy using a combined method of feature selection. At the genetic level, the study revealed a high expression of genes related to cell regulation functions, and a low expression of genes responsible for Mismatch Repair functions, in patients with high instability.

Keywords: Machine Learning, Deep Learning, Cancer, Genome, Immunotherapy, Microsatellite instability, RNAseq.

Resumo

O cancro é uma das principais causas de morte em todo o mundo, tendo provocado mais de 19 milhões de novos diagnósticos e quase 10 milhões de mortes em 2020. Os métodos mais comuns para combater a doença são a rádio e quimioterapia. No entanto, estes deixam uma marca significativa nos pacientes, com efeitos colaterais como queda de cabelo, redução do apetite, fadiga ou diarreia. A imunoterapia é um novo método que está a revolucionar o tratamento do cancro. Apesar de estar numa fase inicial de aplicação, a imunoterapia tem-se mostrado um dos métodos mais promissores para tratar o cancro, permitindo que os pacientes mantenham uma melhor qualidade de vida durante o tratamento e tenham maior expectativa de vida.

O problema abordado é a previsão da Instabilidade de Microssatélite, um biomarcador essencial com uma promissora capacidade de prever a resposta do sistema imunológico. Este dado levou à procura de genes competentes no combate ao cancro no sistema imunológico do paciente. Para encontrar uma resposta para este problema, serão utilizados dados de RNA sequencial para extrair assinaturas expressão genética e mutações, o que permitirá a estratificação de pacientes com cancro para definir melhores planos de tratamento.

Usando dados do TCGA, três diferentes abordagens foram desenvolvidas para prever a instabilidade de microssatélite, com vários métodos de seleção de *features* testados nos algoritmos *Multi-Layer Perceptron*, *Random Forest* e *K-Nearest Neighbors*. O objetivo principal foi a compreensão da capacidade dos diferentes modelos de prever a instabilidade e como cada método seleciona os genes mais relevantes para fazer essa previsão. O estudo teve também como objetivo confirmar a capacidade do RNA sequencial em prever a instabilidade e a comparação dos modelos de *Deep Learning* com outros modelos de *Machine Learning*.

O estudo concluiu que o *Multi-Layer Perceptron* tem uma melhor capacidade de usar dados de RNA sequencial para prever a instabilidade, com a abordagem que uniu pacientes com baixa instabilidade e estabilidade nos seus microssatélites, no adenocarcinoma do colon, a obter um desempenho de 98.44% de AUC e 92.67 % de eficácia, usando um método combinado de seleção de *features*. A nível genético, o estudo revelou uma alta expressividade de genes relacionados com as funções de regulação celular e uma baixa expressividade de genes responsáveis pelas funções de correção de emparelhamento de bases, em pacientes com alta instabilidade.

Palavras-Chave: *Machine Learning*, *Deep Learning*, Cancro, Genoma, Imunoterapia, Instabilidade de Microssatélite, RNA sequencial.

Acknowledgments

With this dissertation, I finish a five-year chapter in my life. I arrived at the University of Porto as a teen, but I leave as a man. In these five years, I pushed through my limits, recognised I have much more strength than I thought I had, and that way, I fought every single day to achieve my main dream: to give my family a better life.

To my supervisors for all the help and advice during this study and for guiding me to follow the right path all the time.

To the friends I have made at Feup, those with me from day 1 to day last. From all the classes to many nights of study and hard work, parties, and so many moments we will not forget.

To the ones that stood with me during the pandemic, to the 'Dysfunctional Family' for all the laughs and the good moments that kept me hopeful for a better tomorrow when we were closed at home. I do not forget everything you did for me.

To IEEE and all that Institution represents to me. It made me achieve my personal best, brought me new friends and challenges, and I hope someday I can return everything that it gave me.

To my family, my parents and my brother. They have been with me throughout the journey, providing everything I needed to become the person I am today. I am very thankful to them, and I hope I can give them what they deserve in the future.

I lived five years at Porto. Being alone in a new city was not easy, but that made me grow. I felt like a stranger in the town for a long time, but that changed when I found that special person. After four years, she was the one that made it feel like home.

To all the mentioned above, Thank you for being part of this. It was a hell of a ride!

Miguel

The best portion of your life
will be the small, nameless moments
you spend smiling with someone
who matters to you.”

Ritu Ghatourey

Contents

1 Introduction **1**

1.1 Motivation . . . 2

1.2 Objectives . . . 2

1.3 Contributions . . . 3

1.4 Document Structure . . . 3

2 From Cancer to Microsatellite Instability **4**

2.1 Gene Mutation . . . 4

2.2 Intratumour Heterogeneity . . . 6

2.3 Microsatellite instability . . . 7

2.4 RNA sequencing . . . 7

2.5 Therapies . . . 10

2.6 Summary . . . 10

3 Literature Review **11**

3.1 Genomics, Intratumor Heterogeneity and Microsatellite Instability Detection . . . 11

3.2 Predictive Models using RNAseq . . . 13

3.3 Discussion . . . 15

3.4 Summary . . . 15

4 Methodology **16**

4.1 Data Description . . . 16

4.2 Data Preparation . . . 18

4.3 Feature Selection . . . 18

4.4 Architectural Design . . . 19

4.5 Summary . . . 20

5 Results **21**

5.1 Microsatellite Instability Prediction Results Analysis . . . 21

5.1.1 Multiclass Approach . . . 21

5.1.2 Binary Approach: Merging MSI-L with MSI-H . . . 23

5.1.3 Binary Approach: Merging MSI-L with MSS . . . 23

5.2 Genetic Selection Results Analysis . . . 25

5.2.1 MRMR . . . 28

5.2.2 ANOVA . . . 28

5.2.3 Dimensionality Reduction and MRMR . . . 30

5.2.4 Dimensionality Reduction and ANOVA . . . 32

5.2.5 15 Feature Set with T-Test . . . 32

5.3 Discussion 32
5.4 Summary 35

6 Conclusions 36

References 38

A Results Table 42

B Hyperparameter Tuning 46

C Heatmaps of the Influential Genes 49

List of Figures

2.1 Gene Expression Change. Retired from [16]. 5
2.2 Genetic factors of intratumour heterogeneity development. Adapted from [8]. . . 6
2.3 Behaviour of proficient (A) and deficient (B) MMR. Adapted from [6]. 8
2.4 RNA sequencing experiment. Retired from [38]. 9

4.1 Diagram of the architectural design of the study. 17

5.1 Heatmap of the most influential genes selected by MRMR with the 3class STAD dataset. 28
5.2 Heatmap of the most influential genes selected by MRMR with the 2class UCEC dataset. 29
5.3 Heatmap of the most influential genes selected by ANOVA with the 3class COAD dataset. 29
5.4 Heatmap of the most influential genes selected by ANOVA with the 2class COAD dataset. 30
5.5 Heatmap of the most influential genes selected by DR + MRMR with the 2 class COAD dataset. 31
5.6 Heatmap of the most influential genes selected by DR + ANOVA in the 2 class STAD dataset. 32
5.7 Heatmap of the 15 genes selected by the 15-feature set with the 3 class UCEC dataset. 33
5.8 Heatmap of the 15 genes selected by the 15-feature set with the 2 class UCEC dataset. 33

C.1 Heatmap of the most influential genes selected by DR + ANOVA with the 2-class COAD dataset. 49
C.2 Heatmap of the most influential genes selected by MRMR with the 2-class COAD dataset. 50
C.3 Heatmap of the most influential genes selected by MRMR with the 3-class COAD dataset. 50
C.4 Heatmap of the 15 genes selected by the 15-feature set with the 2-class COAD dataset. 51
C.5 Heatmap of the 15 genes selected by the 15-feature set with the 3-class COAD dataset. 51
C.6 Heatmap of the most influential genes selected by DR + MRMR with the 2-class STAD dataset. 52
C.7 Heatmap of the most influential genes selected by MRMR with the 2-class STAD dataset. 52

C.8 Heatmap of the most influential genes selected by ANOVA with the 3-class STAD dataset. 53
C.9 Heatmap of the most influential genes selected by ANOVA with the 3-class STAD dataset. 53
C.10 Heatmap of the 15 genes selected by the 15-feature set with the 2-class STAD dataset. 54
C.11 Heatmap of the 15 genes selected by the 15-feature set with the 3-class STAD dataset. 54
C.12 Heatmap of the most influential genes selected by DR + MRMR with the 2-class UCEC dataset. 55
C.13 Heatmap of the most influential genes selected by DR + ANOVA with the 2-class UCEC dataset. 55
C.14 Heatmap of the most influential genes selected by MRMR with the 3-class UCEC dataset. 56
C.15 Heatmap of the most influential genes selected by ANOVA with the 2-class UCEC dataset. 57
C.16 Heatmap of the most influential genes selected by ANOVA with the 3-class UCEC dataset. 58

List of Tables

4.1 Number of samples and features per dataset 17
4.2 Number of samples of each label per dataset. 18
4.3 List of values used to optimize the different algorithms. 20

5.1 Results of the COAD Dataset in the Multiclass approach. 22
5.2 Results of the STAD Dataset in the Multiclass approach. 22
5.3 Results of the UCEC Dataset in the Multiclass approach. 23
5.4 Results of the COAD dataset in the binary classification approach that merges MSI-L with MSI-H. 24
5.5 Results of the STAD dataset in the binary classification approach that merges MSI-L with MSI-H. 24
5.6 Results of the UCEC dataset in the binary classification approach that merges MSI-L with MSI-H. 25
5.7 Results of the COAD dataset in the binary classification approach that merges MSI-L with MSS. 26
5.8 Results of the STAD dataset in the binary classification approach that merges MSI-L with MSS. 26
5.9 Results of the UCEC dataset in the binary classification approach that merges MSI-L with MSS. 27
5.10 Selected experiments for feature selection analysis. 27

A.1 List of the all the results for the experiments with COAD dataset. 43
A.2 List of the all the results for the experiments with STAD dataset. 44
A.3 List of the all the results for the experiments with UCEC dataset. 45

B.1 List of the best parameters for each approach in the Random Forest models. . . . 46
B.2 List of the best parameters for each approach in the MLP models. 47
B.3 List of the best parameters for each approach in the KNN models. 48

Abbreviations

AI	Artificial Intelligence
ANOVA	Analysis of Variance
AUC	Area Under the Curve
cDNA	complementar Deoxyribonucleic Acid
CE	Capillary Electrophoresis
COAD	Colon Adenocarcinoma
DL	Deep Learning
DNA	Deoxyribonucleic Acid
KNN	K-Nearest Neighbours
LOH	Loss of Heterozygosity
ML	Machine Learning
MLP	Multi-Layer Perceptron
MMR	Mismatch Repair
MRMR	Max-relevancy and Min-redundancy
MS	Microsatellite
MSI	Microsatellite instability
MSI-L	Low Microsatellite instability
MSI-H	High Microsatellite instability
MSS	Microsatellite Stability
PCR	Polymerase Chain Reaction
RNA	Ribonucleic Acid
RNAseq	RNA sequencing
ROC	Receiving Operator Characteristic
STAD	Stomach Adenocarcinoma
SVM	Support Vector Machine
TCGA	The Cancer Genome Atlas
UCEC	Uterus corpus endometrial cancer

Chapter 1

Introduction

In 2020, more than 19 million people were diagnosed with cancer, and the number of deaths is close to 10 million, confirming cancer as a leading cause of death worldwide [36]. The most common methods to battle the disease are radio and chemotherapy, but they leave a big mark on the patient. Although most types of radiotherapy are not invasive, the side effects are similar to invasive chemotherapy and are described as hair loss, appetite change, fatigue, or diarrhoea, for example [4].

Female breast cancer and lung cancer are the most common types of cancer, with more than 2 million new cases in 2020, in a percentage of 11.7% and 11.4%, respectively. Besides being second in the number of new cases, lung cancer is the most fatal, responsible for almost 1.8 million deaths in 2020. Female breast cancer appears at fourth in recent deaths, with liver and stomach cancer having more than 750 thousand deaths in the same year [36].

Early detection is essential, with different survival rates associated with different phases in which cancer is detected. Of course, patients with cancers detected in stage I have much more probability of survival than patients with cancers detected in stage IV [10]. For some cancers, even a four-week difference in detection can be crucial for patient survival and success of treatments [9]. Instead of detecting to treat cancer, the prevention of the disease in any of its types can be the solution we need to give a step forward in the next years. Having healthy habits is one way to reduce the probability of cancer, but screening to detect the disease in early stages has helped reduce the death rates of several types of cancer. The technological advancements brought new models to register early biomarkers, with DNA, RNA or immune cells playing important roles to detect malignant cells before they spread. Genomic analyses have provided good descriptions of cancer, and so it seems that there is an enormous potential to improve cancer prevention with an early and accurate stratification of patients [5].

1.1 Motivation

Cancer is defined as a genomic disease characterised by instability in the genome with an accumulation of various point mutations. The immune system appears vigilant, trying to monitor the tumours, with infiltration of immune cells in the tumour micro-environment [41]. Based on these definitions, Immunotherapy arises then as a new method that is revolutionising the cancer treatment, with the goal of reviving the suppressed immune system, boosting the natural defences, to be able to fight the tumour cells and kill cancer [37, 41]. Besides being in a very early stage of development, immunotherapy is establishing itself as one of the most promising methods to treat cancer, while allowing the patients to maintain a better quality of life during the treatment compared to radio or chemotherapy. The life expectancy of patients submitted to immunotherapy is also higher [7].

Several types of Immunotherapy have been developed in the last 130 years, but only recently has this method become an important reference for researchers. The most fundamental types of immunotherapy include Oncolytic virus therapies, cancer vaccines, cytokine therapies, adoptive cell transfer and immune checkpoint inhibitors [41]. The emergence of Immunotherapy as an effective method to treat cancer is much given to T-Cells, which have a powerful tumour-killing capability. However, not all patients can benefit from them, which made researchers look for the benefits of other types of cells, such as B-Cells, Myeloid Cells and NK cells [41].

In this early phase of immunotherapy, numerous challenges are still waiting for an answer, which is what researchers are focusing on now. Machine Learning (ML) is a methodology that is being applied to discover new genetic features and relevant information that may strengthen this therapy, allowing Data Scientists to help solve a real-world problem. ML allows scientists to predict patients' response to immunotherapy by analysing distinct biomarkers that potentiate a better response to this treatment, such as Microsatellite Instability (MSI). This can increase the chance of patients' survival, leading to personalised approaches for each patient or stratified group of patients [26]. Artificial Intelligence (AI) can increase the accuracy of cancer treatment, with immunotherapy taking advantage of AI methodologies to treat cancer successfully. However, there are still numerous challenges guaranteeing the best healthcare quality and patient safety [42].

1.2 Objectives

This dissertation investigates ML's application to predict MSI recurring to RNAseq data. Despite there are some studies about MSI prediction and its relation to immunotherapy, the novel technique includes the use of RNAseq data, which has not been explored much before. This study aims to identify the competent genes of the patient's immune system to predict MS condition and stratify the patients. The analysis of the results will later allow the comprehension of which models are better at prediction with RNAseq data and to confirm with the literature which genes may impact the patients' condition.

1.3 Contributions

This dissertation will include the following contributions:

- Comparison of several ML models to predict MSI using RNAseq data
- Development of three different approaches for each model, a multiclass one and two binaries
- Explanation of the best approach selection from the several models
- Identification of the competent genes in the patient's immune system and how they are represented in each MS condition.
- Understand the possibility of detection of MSI with RNAseq data

1.4 Document Structure

The structure of this document is divided into six chapters. This first chapter explains the motivation, the objectives, and the expected contributions of the dissertation. The second chapter clarifies the medical concepts required to understand the problem. The third chapter analyses the state of the art regarding the problem, explaining what has been done and what gaps exist to explore. The fourth chapter describes how the data was prepared, the feature selection methods used and the selected architectural designs. The fifth chapter exhibits the obtained results and its discussion. Lastly, the sixth chapter specifies the conclusions of this work.

Chapter 2

From Cancer to Microsatellite Instability

Cancer is a disease characterised by an uncontrollable growth of malignant body cells. Usually, when cells are damaged or get old, they die, and new cells take their place. When these abnormal cells begin to grow and multiply, they can form tumours, in a process that can start anywhere in the human body [16, 33]. The American Cancer Society defines a tumour as a lump of growth. Some tumours are cancerous and are called malignant, while others are not and are called benign. Malignant tumours are composed of cancer cells that have specific properties [33].

There are many differences between cancer cells and normal cells. Contrarily to normal cells, cancer cells can grow in the absence of signals and ignore others, such as programmed cell death. They can also invade nearby areas and spread to other parts of the body, a process called metastasis, and communicate with blood vessels to grow towards tumours to guarantee a supply of essential substances. One of the most notable features is that they can mislead the immune system, leading immune cells to protect instead of attacking them. Lastly, they may change their chromosomes, deleting or duplicating certain parts of genome information [16].

Genetic changes in cellular function control cause cancer. These changes can happen due to errors in cell division, DNA damage from harmful substances or inheritance from parents. Each person's cancer is originated from a unique set of genetic changes that commonly affect three different sorts of genes: proto-oncogenes, tumour suppressor genes, and DNA repair genes [16].

2.1 Gene Mutation

Gene Mutations can be defined as changes in the genetic sequences of individuals, specifically in the nucleic acids, which in cellular organisms are the basis of DNA, as we can see in Figure 2.1. These changes can have different consequences, and so we cannot predict if the effect of a mutation is good or not. Positively, they can be responsible for the generation of diversity among organisms

[20]. However, they also increase the risk of disease formation since there is no restriction to the occurrence of DNA sequence changes [39].

Each somatic mutation in a cancer cell genome can be classified between driver and passenger mutations according to its consequences for cancer development. Driver mutations have been positively selected during cancer evolution and grant growth advantage on the cells that carry them. In contrast, passenger mutations do not confer growth advantage, but they were present in the ancestor of that cancer cell when one of the driver mutations was acquired [35].

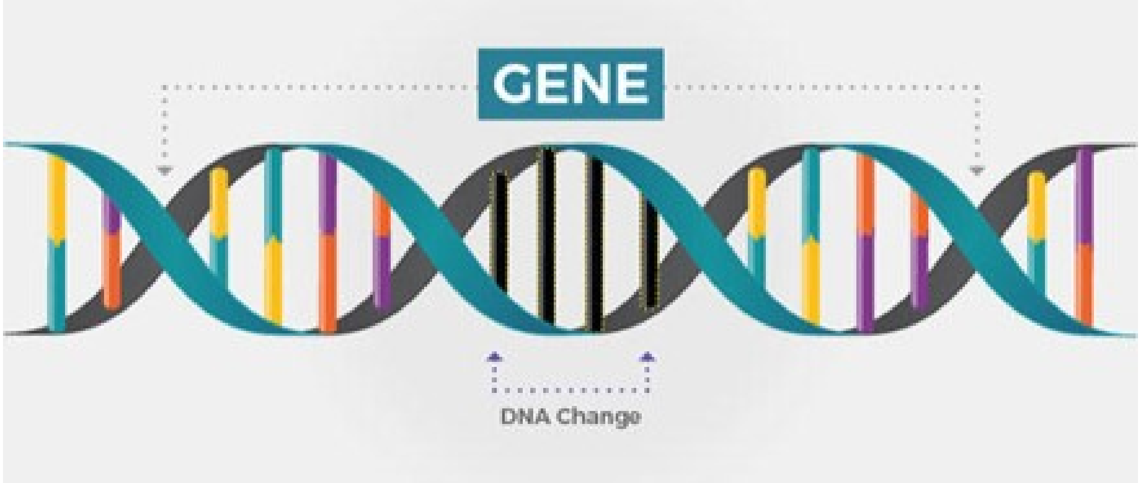

Figure 2.1: Gene Expression Change. Black bases represent changed pairs, while blue-yellow and orange-purple represent A-T and C-G normal pairs. Retired from [16].

The most basic mutations that can occur are point mutations, which are single-base pair changes in the DNA. A point mutation that changes the amino acid sequence is a nonsynonymous mutation. In contrast, a synonymous mutation, also referred to as silent, does not change the amino acid sequence since various codons encode many amino acids. Insertions and deletions are also other forms of mutations with multiple lengths. They can activate new cellular functions or delete the normal ones. Insertions or deletions of one or two base pairs will cause a frameshift in the DNA because proteins are codified by a three-base pair codon [20, 39].

With DNA having two copies of the same gene [21], a deletion that affects only one of the copies happens in regions of LOH. LOH is usually the first strike to inactivate a tumour suppressor gene in sporadic cancers. The second strike is the one that occurs in that gene. Deletions that affect both copies of a gene are called homozygous deletions and are sometimes observed in the cancer genome, being a signal that a tumour suppressor gene was located in the lost region [39].

Usually, cancer will not be caused by one mutation, but it will result from an accumulation of mutations during the patient's lifetime. That is also the reason why cancer is most probable in older people [35].

2.2 Intratumour Heterogeneity

A tumour can be described with intratumor heterogeneity when it has distinct tumour cell populations within the same tumour specimens. These differences can be found in the molecular and phenotypical profiles and are very common in malignant tumours [40]. Intratumor heterogeneity manifests itself in the absence of uniformity of morphological structures or genotypic status and the variable expression of different markers by distinct tumour cell groups within the same tumour [8].

The arising of intratumor heterogeneity is probably a source for adaption of the tumour to alterations in the micro-environmental conditions, playing an essential role in distinct forms of tumour progression. Growth and invasion of a primary tumour and lymphogenic and hematogenic metastasis are seen as principal factors defining the development of tumours. These factors allow the progression with the capacity to maintain oncogenic potential, cell survival under conditions of the non-static micro-environment, and tumour resistance to drug therapy [8].

The factors that lead to the development of intratumor heterogeneity can be divided between genetic and non-genetic [8]. This study will focus on the genetic ones, using data generated from the patient's genetic expression. As represented in Fig. 2.2, genetic factors include chromosomal instability, gene mutations and microsatellite instability. All of them lead to high genome instability, which leads to an increase of clonal diversity and genetic, epigenetic, and phenotypic heterogeneity [8].

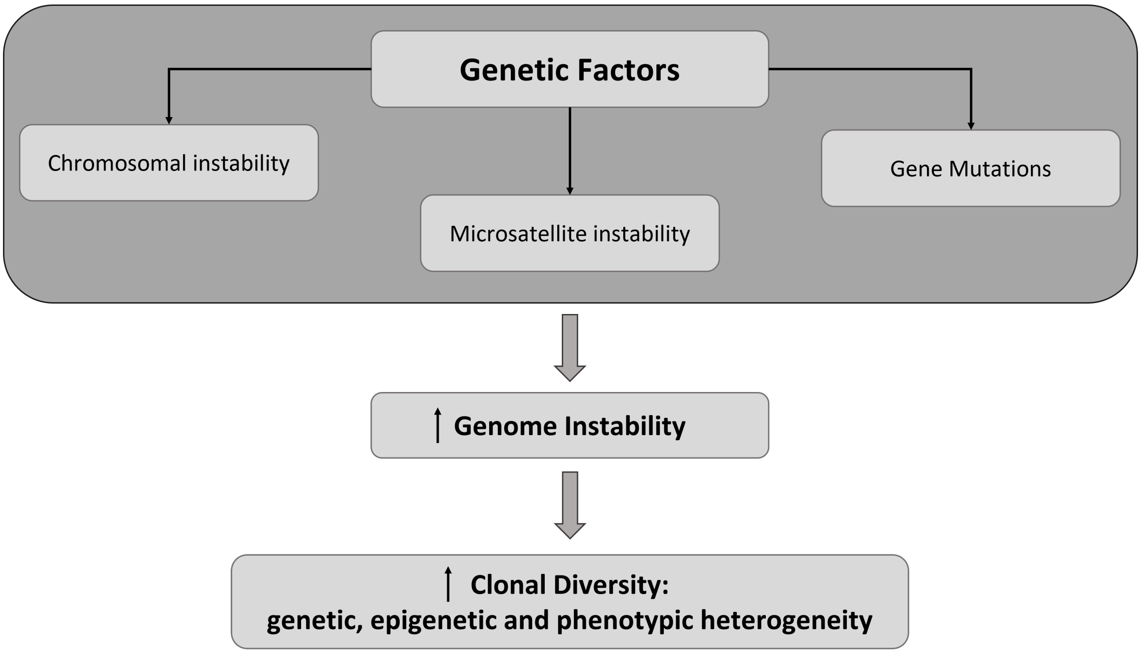

Figure 2.2: Genetic factors of intratumour heterogeneity development. Adapted from [8].

Intratumor heterogeneity is one of the main determinants of therapeutic resistance and treatment failure. In patients with metastatic diseases, it is also one of the main reasons for low overall survival. The challenges for precise treatments to different patients are not being overcome because the distinct cell populations have other characteristics and react differently to the therapies. This means that not all the cancer cells will respond positively to the treatment, leaving a significant chance of survival to some of the populations, which increases the risk of recurrence in the future [40].

2.3 Microsatellite instability

Microsatellite (MS) consists of repeated sequences of 1 to 6 nucleotides, tandemly distributed in a series of 15 to 65 nucleotides of small satellite DNA, mainly situated near the ends of chromosomes. Each MS-specific place comprises two parts: the central core and the peripheral flanks. The variation in the number of core repeating units originates the specificity of MS [18].

Mismatch Repair (MMR) is the normal tissue DNA repair system that can correct errors in the process of DNA replication. The possibility of gene mutation increases with the lack of MMR genes in tumour cells or errors in the process of replication repair [18]. Figure 2.3 illustrates how both proficient and deficient MMR work in the correction of errors in DNA replication. In (A), MSH2, MSH6, PSM21, and MLH1 proteins work successfully to repair the incorrect base, while in (B), those proteins can't cooperate and the result is a defective DNA.

With the definitions of MS and MMR, microsatellite instability (MSI) can be then defined as a hypermutable phenotype caused by the loss of DNA MMR activity [1].

Distinct methods are being used to detect microsatellite instability, such as next-generation sequencing, fluorescent multiplex PCR and CE, immunohistochemistry, single-molecule molecular inversion probes and MSI calculation method. From this list of methods, the fluorescent multiplex arises as to the gold standard, reaching an accuracy of 100%, but only obtaining MSI results [18].

MSI has been detected in 15% of all colorectal cancers, characterising them with distinctive features and giving them a better prognosis than colorectal tumours without MSI. The discovery of MSI in this specific type of tumours is helping the application of personalised treatments to patients, which shows the importance of determining this biomarker [1].

2.4 RNA sequencing

The transcriptome comprehends the complete transcripts and their quantity in a cell for a particular developmental phase or physiological condition. It's essential for the interpretation of the functional elements of the genome to reveal the molecular components of cells and tissues and to understand development and disease. Within the main goals of the transcriptome, we can highlight the determination of the transcriptional structure of genes and the quantification of changing expression levels in each transcript [38].

RNA sequencing (RNAseq) is an approach to transcriptome profiling that uses deep-sequencing technologies. The procedure converts a DNA population, total or fractionated, to a library of cDNA fragments with adapters attached to one or both ends. Each molecule, with or without

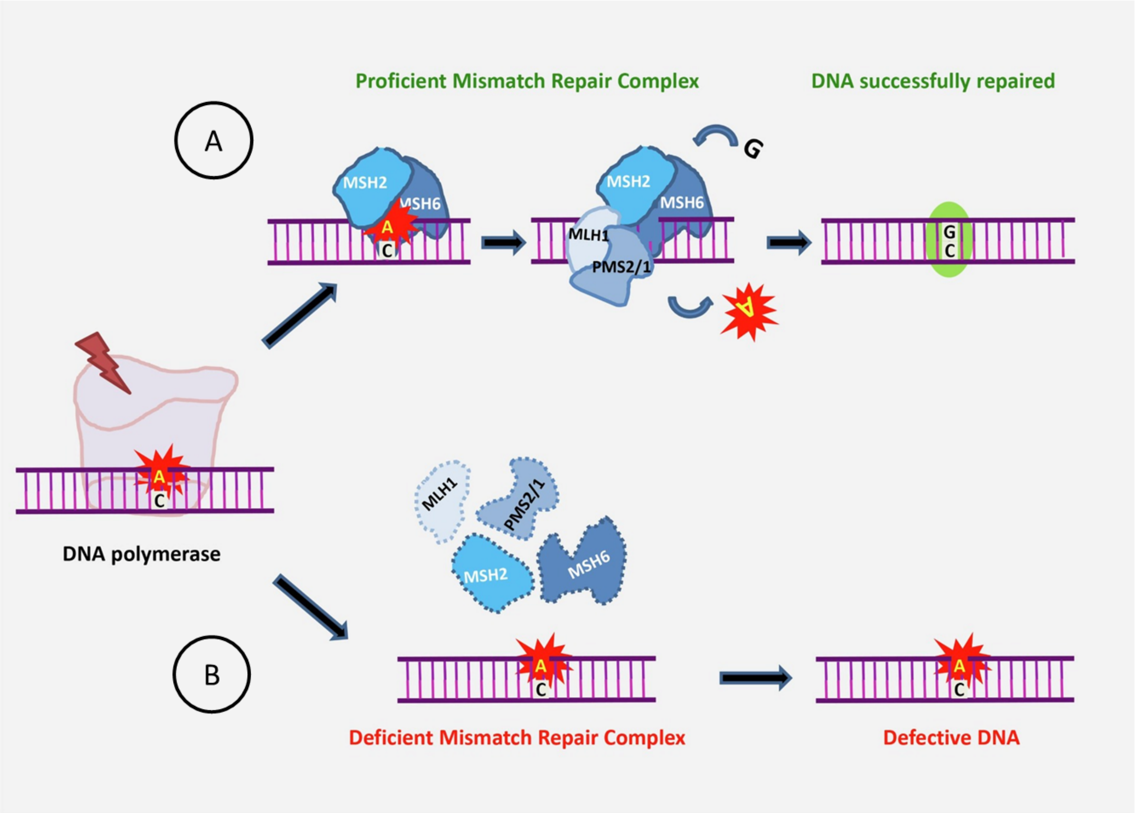
Proficient Mismatch Repair Complex
DNA successfully repaired
A
MSH2
MSH6
A
C
MSH2
MSH6
MLH1
PMS2/1
G
A
G
C
DNA polymerase
A
C
MLH1
PMS2/1
MSH2
MSH6
B
A
C
A
C
Deficient Mismatch Repair Complex
Defective DNA

Figure 2.3: Behaviour of proficient (A) and deficient (B) MMR. Adapted from [6].

amplification, is then translated in a high-throughput mode to obtain short sequences from one or both ends. After sequencing, the resulting reads are lined up to a reference genome or assembled again without the genomic sequence to produce a genome-scale transcription map. This map can comprehend both the transcriptional structure and the expression level of each gene, or just one of them [38] This process is illustrated in the Figure 2.4 below.

RNAseq offers a large number of advantages compared to other existing technologies, as Wang et al. explain in Table 1 of [38]. With RNAseq, the exact location of transcriptome boundaries can be revealed to a single-base resolution and sequence variations in transcribed regions. RNAseq has a high throughput, which is helpful to study complex transcriptomes, a very low background noise, and can detect transcripts over a big range of expression levels (greater than 9,000-fold). It's highly accurate in quantifying expression levels, showing high levels of reproducibility for technical and biological replicates while requiring less RNA amount than the other methods. Concerning practical issues, it has a relatively low cost for mapping transcriptomes of large genomes. With these benefits, RNAseq is the first sequencing-based method allowing all the transcriptome to be surveyed in a very high throughput quantitatively while offering a single-base resolution for annotation and gene expression levels at the scale of the genome, with a much lower cost than any other method [38].

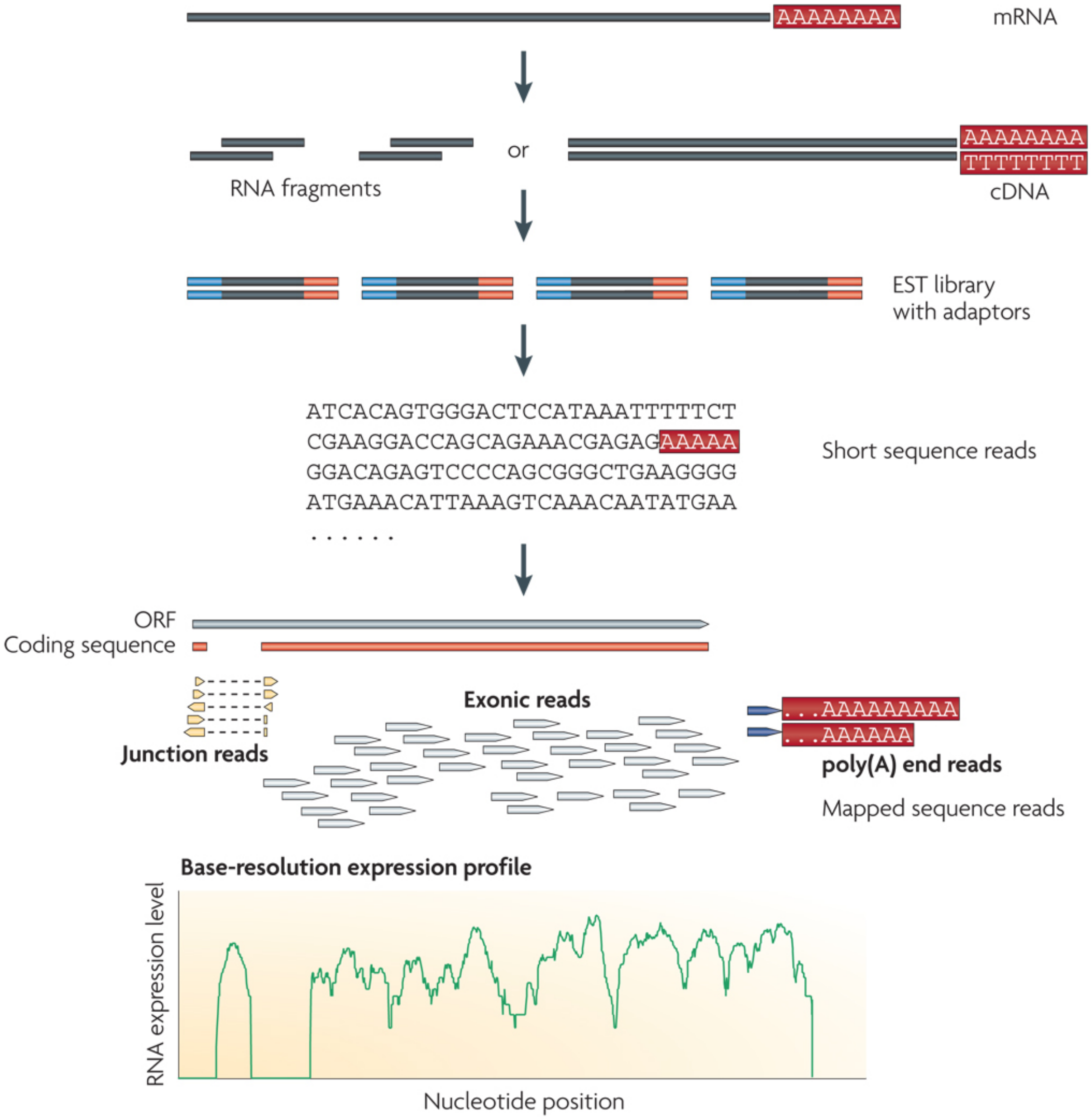

Figure 2.4: RNA sequencing experiment. mRNA data is converted in RNA fragmens or cDNA. Sequencing adaptors are added to each cDNA fragment and the resulting sequence reads are aligned with reference genome. They can be classified as exonic reads, junction reads and poly(A) end-reads. These three types are then used to generate a base-resolution expression profile for each gene. Retired from [38].

2.5 Therapies

Several methods to treat cancer have been developed over the years. Their prescription depends on the type of cancer and the stage of development it is, with the majority of the people receiving not one, but a combination of treatments [13]. Commonly, radio and chemotherapy are the most known methods, often allied with surgery. Still, immunotherapy is emerging as a less invasive treatment with better results when compared with the previous two, revolutionising the way of treating cancer worldwide [7].

Radiation therapy or **radiotherapy** is a treatment consisting of the use of high doses of radiation to kill cancer cells and shrink tumours. It aims to kill or slow the growth of cancer cells by damaging their DNA. This therapy does not kill cancer cells immediately. There is a need for several days or weeks of treatment to damage sufficiently the DNA to provoke their death [32, 15].

Chemotherapy is a treatment consisting of the use of drugs to kill cancer cells. It aims to slow or stop the growth of cancer cells [14]. Several types of drugs can be used, each one with its advantages and effects, as referred in [31]. It can be used to cure cancer, to control the disease and in palliative care, helping the person to feel better and live longer [30].

Both these treatments leave a significant mark in cancer patients, with side effects that affect their quality of life in several ways. Besides varying from treatment to treatment and patient to patient, the most common effects are appetite loss, fatigue, diarrhoea, hair loss, nausea or vomiting. With this in mind, it's essential to improve the ways of treating cancer to guarantee to the patients the best possible conditions for living [4].

Immunotherapy is an emerging method that is gaining much interest in the scientific community, bringing great results and fewer side effects for patients when compared to radio or chemotherapy. Immunotherapy aims to revive the suppressed immune system, boosting the patient's natural defences to attack tumour cells and kill cancer. There are different types of immunotherapy, able to reactivate the benefits of different types of cells, but all of them are contributing to improving cancer patients' life expectancy [7, 37].

2.6 Summary

Cancer is a genomic disease caused by an accumulation of DNA mutations that generates cells with distinct characteristics that can mislead the normal function. MSI can indicate cancer by the loss of DNA MMR activity. RNAseq is a recent approach to transcriptome profiling that can quantify the gene expression levels accurately. Immunotherapy is an emerging method to treat cancer that uses the patient's immune system to battle the disease.

Chapter 3

Literature Review

To better approach the problem of this work, two different areas in the literature were studied: Genomics, Intratumor Heterogeneity and Microsatellite Instability with RNAseq Statistical Analysis, and Predictive Models using RNAseq. The following sections comprehend the most relevant studies.

3.1 Genomics, Intratumor Heterogeneity and Microsatellite Instability Detection

Different researches were found to study the biological part of this problem, characterising the human body genome, intratumor heterogeneity, and MSI. The analysed studies are from different databases, but both Nature and Cell represent essential fonts of information in the area. The analysed studies were found in Google Scholar[1], employing the query *("microsatellite instability detection" AND "artificial intelligence")*.

In 2020, Li et al. [18] studied the relation between MSI and immunotherapy. Four methods are highlighted in the detection of the condition, with fluorescent multiplex PCR and CE having the golden standard of 100% accuracy, better than the 95.8% of Single-molecule molecular inversion probes, the 92-94.6% of Next-Generation sequencing and the 89-95% of Immunohistochemistry. The analysis of mutation characteristics also showed some recurrent features, indicating that the specific tumour environment conduces to MSI events. MSI occurs mainly in ion-binding genes in gastric adenocarcinoma, with tumour suppressor genes ACVR2A and RNF being the most common targets of mutations in tumours with high MSI. MSI has been found in several cancer types, e.g. gastric, breast, prostate, ovarian, and endometrial, which shows how crucial it can be as an indicator. The author concludes that MSI leads the tumour to be drug-resistant and is an effective positive immunotherapy predictor.

[1] https://scholar.google.pt/, last accessed on 25/08/2022

In 2018, Matak et al. [22] characterised stochastic phenotype switching as a mechanism leading to intratumor heterogeneity. The author established a primary tumour cell culture from tissue from a surgically resected sarcomatoid cholangiocarcinoma of the liver. Using RNAseq, it was investigated the presence of different types of keratin-7 cells and their capacity to change their transcriptional profile, switching between different phenotypes. While keratin-7-negative cells are stable, keratin-7-positive were able to change their phenotype, generating unstable and heterogeneous populations. A relation between the loss of keratin-7 expression and tumour formation was found, with keratin-7-negative cells indicating increased tumorigenic potential.

Hugo et al. [12] studied the genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma. The authors analysed whole-exome sequences in 38 pre-treatment melanoma tumours, from responsive patients with complete or partial responses to anti-PD-1 and non-responsive patients with progressive disease. The transcriptomes were analysed, recurring to RNAseq data. Responding tumours had harboured more non-synonymous single nucleotide variations. The main conclusion from this work associated a higher mutational load significantly with better survival in melanoma patients, but it was not predictive of response to anti-PD-1 therapy.

The TCGA Research Network also performed some exciting studies that became more relevant with the definition of TCGA as our central database. In 2014, the study in [25] analysed the molecular characterisation of gastric adenocarcinoma using six different molecular platforms, one of them being RNAseq. They proposed a molecular classification of this type of cancer, with one of the groups being relative to unstable microsatellite tumours. A HotNet analysis of mutated genes in the MSI group showed alterations in primary histocompatibility complex class I genes, B2M and HLA-B. B2M mutations are common in colorectal cancers and melanomas. Usually, they result in loss of expression of HLA class 1 complex, suggesting hypermutated tumours would benefit from this by reducing antigen presentation to the immune system. An elevated C to T base transition rate at CpG dinucleotides was also observed. RNAseq data showed 11 fusions between CLDN18 and ARHGAP26 or ARHGAP6, contributing to the invasive phenotype of diffuse gastric cancer. Mutations were also found in PIK3CA, ERBB3, ERBB2 and EGFR, while mutations in BRAF and V600E, common in MSI in colorectal cancer, were not found.

Two years before, they also studied the molecular characterisation of colon and rectal cancer in humans [24]. Their work identified MSI in 75% of the hypermutated colorectal carcinomas and MMR gene mutations in the other 25%. In the proceedings, the authors have divided the samples between hypermutated, if the mutation rate was above 12 per 10^6 bases, and non-hypermutated, if below 8. Within the hypermutated tumours, 77% had high levels of MSI. In contrast, 23% did not have that high level but had somatic mutations in at least one MMR gene or polymerase ε aberrations. They also found 15 recurrently mutated genes in the cases with hypermutation, with ACVR2A, APC, TGFBR2, BRAF, MSH3 and MSH6 having mutation values above 40%. The hypermutated tumours with silencing of MLH1 and high levels of MSI showed additional differences in the mutational profile.

Recently, Hildebrand et al. [11] studied the application of AI to detect MSI and predict the response to immunotherapy in colorectal cancer. Their article focused on reviewing several other studies from different authors with several types of data, but none of them focused on using RNAseq data. The histological features' analysis was the prediction's primary focus, but it lacks any new approach since it was just a review article. It gives a clear idea of what can be achieved with histology, and it is helpful to understand if RNAseq data can be promising for detecting MSI.

3.2 Predictive Models using RNAseq

The second group of selected studies focuses on developing predictive ML models using RNAseq data with different goals. The search employed the query *("Predict microsatellite instability") AND ("using RNAseq data")* in Google Scholar[2] and IEEE Xplore[3]. This section provides information about the model development, feature extraction, parameter tuning and accuracy scores of previous work that can help us to generate the desired models.

The study from Li et al. [19] was one of the main findings of our search, becoming a fundamental research to help lead the direction of this work. The authors focused their work on the prediction of MSI, using mRNA, RNAseq and microarray gene expression data. They selected a 15-feature set using the t-score metric, which will also be used in this study. In their models, just a binary classification approach was tested, merging MSI-L and MSS classes, justifying it with the genetic proximity between them. The only tested model was the KNN, producing results of 97% AUC and accuracy in STAD, 97% AUC and 96% accuracy in COAD, and 93% AUC and 90% accuracy in UCEC.

Pacinkova and Popovici [27] produced the first gene signature to predict MSI with two different types of data, RNAseq and microarrays. The data used to develop the model was only from colon cancer patients. However, the tests performed well in data from gastric and endometrial cancer patients, suggesting a typical pattern between the expression of distinct cancer in the 25-gene signature produced.

Danaher et al. [3] generated an experience to predict MSI in COAD, STAD and UCEC in a 14-gene feature set, focused on mismatch repair genes. The results showed a high prediction accuracy, sensitivity and specificity. However, the authors state that their work might have some limitations since it was only developed under TCGA datasets. Their main conclusion was the possible use of their work to create a combined approach to measure tumour antigenicity and the presence of a suppressed adaptive immune response in a single platform.

In 2021, Sorokin et al. [34] aimed to validate three different studies, [19], [27] and [3] on MSI prediction using RNAseq data. Their motivation was that those studies were never validated in independent databases outside TCGA. The authors executed the available code from the analysed studies with independent datasets and confirmed the effectiveness of the work from Li and Danaher in [19], and [3], respectively. On the other hand, the model from Pacinkova [27] was only effective in colorectal, oesophagal and uterine cancers. The authors explained that conclusion with the lack of gene ontology term "biological processes" in the gene feature set.

[2] https://scholar.google.pt/, last accessed on 25/08/2022

[3] https://ieeexplore.ieee.org/, last accessed on 25/08/2022

Early this year, Seo et al. [29] developed a single transcriptome predictor of MSI for colorectal cancer called MAP (Microsatellite instability Absolute single sample Predictor). This model was developed with a Random Forest algorithm and a recursive feature elimination method. The results in the tests with colorectal patients' data reached an accuracy of 96%, while the validation with STAD and UCEC reached accuracies of 80% and 75%, respectively.

Cascianelli et al. [2] studied breast cancer classification into distinct subtypes with ML methods and RNAseq, giving an important insight for this dissertation about the efficiency of different methods. The approach used the PAM50 classification as a reference, comparing, for each sample, the relative expression of 50 genes. The authors proved that choosing samples to build the PAM50 reference affects the subtyping classification. Then, a new strategy was proposed, called AWCA, in which, after the PAM50 classification, the average expression is calculated within each subtype. Then, with the mean values for each subtype, a new average is calculated for each gene, resulting in the final reference values of the new method. In the ML phase, a random and balanced training set of 220 samples was extracted from TCGA. Five different methods were applied to classify the cancer samples, with 10-fold stratified cross-validation and hyperparameter grid search being applied. The results showed similar training accuracies, between 84 and 88%, but the test phase showed different scores. Support Vector Machines (SVMs) had 74% testing accuracy, while the training was 84%. Feed Forward Neural Networks had the lowest test accuracy, with 64% being paired with 86% in the training step. Since the training data is balanced, we can assume some overfitting of the algorithms. On the other hand, Multiclass Logistic Regression (MLR) showed the best score overall, with 88% of accuracy in training and 85% in testing. The algorithm was then selected to perform new classifications, focused on different feature spaces, to improve the classifier performance. MLR showed potential for improvement in breast cancer subtypes classification by exploring relevant discriminative genome parts using RNAseq data.

Lapuente-Santana et al. studied the presence of biomarkers to predict cancer patients' response to immune-checkpoint blockers (ICB) therapy, a very promising treatment for several types of cancer [43]. The authors of this work aimed to quantify tumour-infiltrating immune cells, the activity of intracellular signalling and transcription factors, and the extension of intracellular communication using RNAseq data. Then, they applied ML models to understand how these signatures are associated with 14 different transcriptome-based predictors of anticancer immune responses. The dataset included 7550 samples from patients of 18 types of cancer generated by TCGA. The authors used two distinct multi-task ML models, a regularised multi-task linear regression and a Bayesian efficient multiple-kernel learning, with the scores, from the 14 predictors as input data, identifying ten highly correlating scores. Immune cells were identified as biomarkers of immune response with $CD8^+$ T cells being considered the most essential for tumour-cell recognition and killing. TRAIL, JAK-STAT, and NF-kB signalling pathways were positively correlated with the predicted immune responses in all cancer types, indicating an expected good response to ICB therapy. On the other hand, VEGF had the most negative correlation. The authors stratified patients according to their expected reaction to ICB therapy, recurring to the multi-task ML methods and the Estimate Systems Immune Response (EaSIeR) system, with AUC scores between 0.78 and

0.85.

3.3 Discussion

In Section 3.1, we can understand how RNAseq can be used to extract indicators related to specific biological events. In the past years, scientists have focused on applying several statistical methods to the data extracted from the novel approach of RNAseq to find essential features that can help the diagnostic and treatment of cancer and help to save lives. Several advances have been made in genomics, with genetic expression and cellular characteristics being revealed as crucial features in cancer detection and patient response.

The evolution of ML is opening a new chapter in the data analysis field. Several areas are being improved by applying AI methods and models that learn from previous data. Health and biology are not different, and we can see some examples in Section 3.2. Predictive models are increasingly applied to understand behavioural patterns and stratify cancer patients for different therapy applications or responsive analyses.

In the studies on 3.2, it is possible to see how ML algorithms such as KNN and Random Forest produce excellent results in predicting MSI from RNAseq and other data types. However, with promising results, there is still a lot to explore. Genomics is still an area with high research potential, and ML is revolutionising the capacity to make discoveries and validate theories. Inside the ML area, DL seems to be a field that can be analysed, potentially improving the already obtained results and helping to give a new step in cancer patients' treatment.

This group of analysed researches show the importance of the work developed in this dissertation. Despite the different approaches analysed in these studies, none directly compared the results of ML and DL models, which is also a gap this thesis will explore.

3.4 Summary

This chapter explored the literature in genomics, RNAseq and ML. The first section looked at the comprehension of Microsatellite Instability detection, with studies giving a relevant statistic analysis and domain reviews about the problem concepts. The second part was focused on MSI prediction using several distinct models and approaches with RNAseq data. Those allowed us to understand state of the art in this area and define the main goals of this work.

Chapter 4

Methodology

The approach to answering the central questions of this dissertation included four main steps explained in the sections below. Chronologically, in the first stage, we focused on understanding and interpreting the used datasets to extract some knowledge about the data. That step is described in Section 4.1 by explaining the characteristics of the three datasets used in this work. In the second stage, those datasets were adequately prepared to reduce their dimensionality and the effect of their imbalance, as it is demonstrated in Section 4.2.

With still a large number of features after dimensionality reduction, it was necessary to apply some feature selection methods to choose the essential features in MSI prediction as specified in Section 4.3. In the final stage, the ML algorithms were selected and optimised for each dataset in an iterative process guided by the results obtained. Section 4.4 explains the followed methodology in this work. To better understand the followed methodology and the logical sequence of the chapter, figure 4.1 represents the architectural design of this work schematically.

4.1 Data Description

Three datasets were used in this project, all with data from the extensive TCGA database[1]. They comprise RNAseq data with the MS label for each patient, generating datasets with more than 57 thousand features, each corresponding to a different human gene found in the samples of the respective dataset. Regarding the labels, TCGA classifies MS in three different categories: high instability (MSI-H), low instability (MSI-L) and stability (MSS). MSI-H and MSS are the most different ones, while MSI-L is located between both. The number of samples and features in each dataset is exhibited in Table 4.1.

RNAseq data and MS data were in a different sub-database in TCGA. So, it was necessary to merge the various tables to produce the desired datasets. Since both had the patient's code in each sample, the merge was done with that unique identifier, generating the datasets used in this

[1] https://portal.gdc.cancer.gov/, last accessed on 15/04/2022

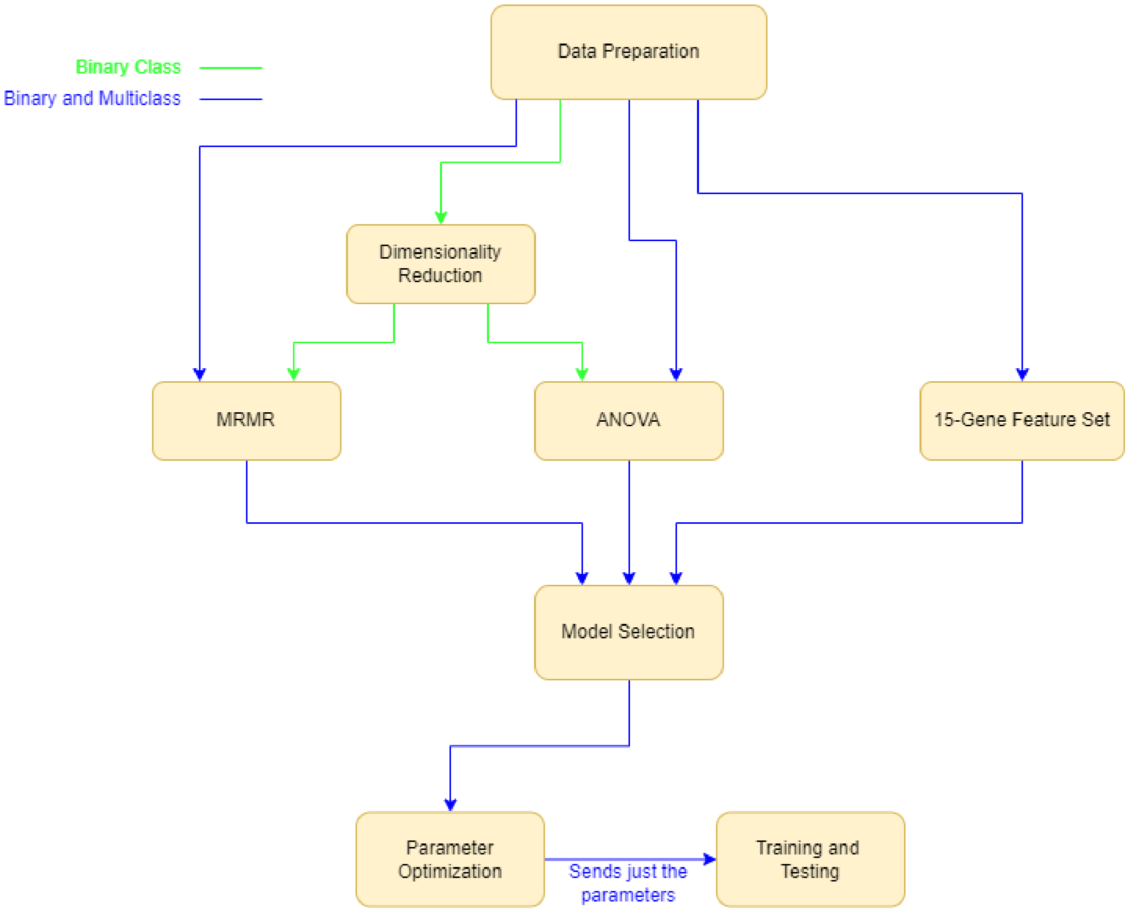

Figure 4.1: Diagram of the architectural design of the study.

study. The datasets have numerical values representing the count of each gene expression for each samples. This allow us to have the expression level of each gene in each patient, potentiating the analysis of which genes stand out in each group of patients.

Table 4.1: Number of samples and features per dataset

Dataset	**Samples**	**Features**
COAD	474	57 137
STAD	407	58 428
UCEC	521	57 763

The **COAD Dataset** has 474 samples, with 57137 features from patients detected with Colon Adenocarcinoma. It has 91 MSI-H samples, 86 MSI-L samples and 297 MSS samples, as it is possible to seen in Table 4.2. This dataset is highly unbalanced, with more than half of its representation in samples from patients with MSS.

The **STAD Dataset** has 407 samples, with 58428 features from Stomach Adenocarcinoma patients. It is even more unbalanced than the previous dataset, with 74 MSI-H samples, 62 MSI-L samples and 271 MSS samples, with these last representing almost 67% of the dataset. Its numbers are shown in Table 4.2.

The **UCEC Dataset** is the largest of these three first datasets, having 521 samples with 57763 features from Uterus Corpus Endometrial Cancer patients. In terms of balance, it is slightly different from the previous two since the number of MSI-H samples is almost four times higher than the number of MSI-L samples: 166 to 45. More than half of the dataset is still from MSS samples, with 310. Table 4.2 exhibits its numbers.

Table 4.2: Number of samples of each label per dataset.

Dataset	MSI-H	MSI-L	MSS	TOTAL
COAD	91	86	297	474
STAD	74	62	271	407
UCEC	166	45	310	521

4.2 Data Preparation

Due to a large number of features in each dataset, the first step was to clean the dataset and eliminate the redundant or non-significant features. Those features with the same value in at least 50% of the samples, and those whose range was less than or equal to five, were discarded since their information was not relevant to the algorithms. Removing these features reduced each dataset to approximately half its original size.

In the binary classification approaches, the next step included another dimensionality reduction that allowed us to make a first extensive selection of features. That step was not performed in the multiclass classification approaches since the method only worked with binary classifications. In this stage, we reduced the feature space from around thirty thousand to about three thousand by calculating the importance of each feature in the dataset and selecting the highest ones until the set of chosen features reached a cumulative importance of 0.99. This method is open source, and it was found on GitHub [17].

The next step was the main Feature Selection stage, but since the available datasets were extensive and it was essential to test different methods, Section 4.3 will explain that process.

4.3 Feature Selection

In the Feature Selection stage, we applied three distinct methods to understand the effect of different genes while predicting the MS condition. The first two were MRMR and ANOVA, while the last used a set of 15 features from a study on the same topic, in which the authors performed a T-Test score to calculate the feature importance [19]. The choice for this last feature set had the goal of comparing the obtained results with the ones from the study and validating those results in this work. With MRMR and ANOVA, we decided to select the 100 best features.

The **MRMR** selects a k-specified number of features by applying a Max-Relevance and Min-Dependency criterion. This technique aims to select the most critical mutually exclusive features in a dataset, which means the relevant ones with less similarity between them [28]. The **Max-Relevance** step generates a set with the most important features by calculating the mutual information value between each feature and each label. Then, the **Min-Redundancy** process will select only mutually exclusive features of the previous set [28].

ANOVA stands for Analysis Of Variance. It works by calculating the variation between the mean of each class for a feature, with the null hypothesis of no significant variation existing. The alternate indicates a variation in the class means, so the features confirming this hypothesis are stored. In the last set of features, the F-Measurement will use a p-value to define the selected features by the method [23].

In the last method, the study from [19] was replicated using the 15-gene set the authors selected in their work. The decision to not perform the t-Test they did to generate the feature set and instead use the set they produced was to understand if a smaller set would produce many different results in other ML models. That way, we could validate their results in the most reliable way possible.

This process was one of the crucial stages of this work and implied the most search since it is not easy to select from large datasets. Identifying the genes with the most impact in the prediction of MS condition was one of the goals of the developed study. With each method working differently, the feature selection would always generate different feature sets from dataset to dataset.

4.4 Architectural Design

After the data preparation phase, the feature selection procedures were tested in each model. The binary classification approaches investigated five procedures. Two of them were composed of the dimensionality reduction phase of [17] combined distinctly with MRMR and ANOVA methods. At the same time, the other three were the simple MRMR, ANOVA and the 15-gene set from [19]. The combined method procedures could not be applied in the multiclass classification approaches, so the tests focused on the other three singular procedures.

Except for the 15-gene feature set, 100 features were selected in the other methods. Some tests were performed to understand the best feature space dimension to select in each procedure, but that value differed from one another and from dataset to dataset. Consequently, the best option seemed to be selecting a fixed number of features for every approach. That would allow the study of the behaviour of a good group of genes in the context of the problem and which would be the most powerful ones in predicting the MS condition in each patient.

In the binary classification models, two different hypotheses were tested. Since each dataset had three different classes, two of them were merged to produce datasets with just two labels. MSI-H and MSS were the most distant ones, while MSI-L was the intermediate class. The choice was then to merge MSI-H with MSI-L in one approach and MSS with MSI-L in another. The literature showed that MSI-L and MSS were the closer classes. However, the two different combinations of that hypothesis would allow us to confirm that fact and check the distance between MSI-L and MSI-H, the two instability classes.

Three different ML models were selected to generate a solution to the proposed problem. First of all, the Random Forest algorithm was selected. The main reason for the option was to gain sensibility to the data and understand how a generic model would work with unbalanced datasets with many features. The DL field was also an area to explore, so the second selected model was a Multi-Layer Perceptron since it has a significant advantage in learning non-linear models. The last chosen model was the KNN, enabling the complete validation of the 15-gene feature set study.

In the optimization phase, a grid search was implemented to find the best combination of parameters and produce the best possible results in each model. Table 4.3 shows the values used for optimization in the three algorithms. Fifty different stratified splits were generated from each dataset to help find those parameters. The parameters were selected from the split with the best score. In an independent process, the training and testing phase was done with a similar split to reduce the overfitting and guarantee the accuracy of the results. In both phases, SMOTE was applied to help balance the datasets and generate more accurate results.

Table 4.3: List of values used to optimize the different algorithms.

Algorithm	**Parameters**	**Values**
Random Forest	n_estimators	100, 150
	criterion	’entropy’, ’gini’
	max_depth	2, 3, 4
	min_samples_split	6, 8, 10
	min_samples_leaf	2, 4
MLP	hidden_layer_sizes	(8,), (8,8), (8,8,8), (16,16), (16,8,16)
	activation	’identity’, ’tanh’
	alpha	0.0001, 0.05
	max_iter	100, 150, 200
	max_fun	10000, 15000, 20000
KNN	n_neighbours	5, 7, 9, 11
	algorithm	’auto’, ’ball_tree’, ’kd_tree’, ’brute’
	leaf_size	3, 5, 7, 10
	p	1, 2

4.5 Summary

Three datasets were used in this work, all part from the TCGA database. Their preparation cleaned the data by eliminating redundant or non-significant features. Three different methods were used to select the most relevant features. Three different algorithms were optimized to train and test the data in several approaches.

Chapter 5

Results

The different tests were performed to understand the possibility of MSI prediction using RNAseq data and to identify the competent genes. Three different approaches were generated to predict MSI, one multiclass and the other two in a binary classification system, to understand if the intermediate class of low microsatellite (MSI-L) instability was distinct from high microsatellite instability (MSI-H) and microsatellite stability (MSS) classes or close to one of them. Five feature selection methods were also tested, two of which were composed of two techniques. The prediction models results is explained in section 5.1, the feature selection method analysis in section 5.2, and its discussion in section 5.3.

5.1 Microsatellite Instability Prediction Results Analysis

Considering the three types of cancer (COAD, STAD and UCEC), the three different classification approaches (one multiclass and two binary), the three different ML algorithms applied (Random Forest, MLP and KNN), and the five different feature selection techniques applied (just three for the multiclass classification approaches), 117 experiments were done in this study. Appendix A shows a table with the results from all those experiences that will be explained in different parts in this section. First, an analysis of the multiclass classification approach is made in subsection 5.1.1, to finish with an analysis of the binary classification approaches in subsections 5.1.2 and 5.1.3.

Each table in this section will have the results presented in percentage, with the mean and standard deviation in two decimal places. Each experiment was fine tuned with a grid of parameters. The set that produced the best result in each of them is presented in Appendix B.

5.1.1 Multiclass Approach

Table 5.1 shows the metrics of the multiclass experiments with the COAD dataset. The MLP model has the best results in all of them except Specificity, where the Random Forest stands. The

Table 5.1: Results of the COAD Dataset in the Multiclass approach.

COAD Dataset

Multiclass Multi-Layer Perceptron

Feature Selection Method	AUC (%)	Balanced Accuracy (%)	Specificity (%)	Precision (%)	Sensitivity (%)	F1 Score (%)
MRMR	**78.38 +/- 3.59**	62.32 +/- 5.30	64.61 +/- 8.30	**69.75 +/- 4.01**	58.65 +/- 4.87	61.49 +/- 4.33
ANOVA	77.86 +/- 2.71	**62.34 +/- 4.84**	63.91 +/- 6.51	69.24 +/- 3.15	**61.26 +/- 5.16**	**63.15 +/- 4.53**
15 Feature Set	76.88 +/- 2.85	61.54 +/- 4.67	**68.62 +/- 7.38**	68.12 +/- 4.78	54.17 +/- 5.67	56.11 +/- 6.64

Multiclass Random Forest

Feature Selection Method	AUC (%)	Balanced Accuracy (%)	Specificity (%)	Precision (%)	Sensitivity (%)	F1 Score (%)
MRMR	**71.33 +/- 3.76**	**55.58 +/- 5.37**	64.76 +/- 12.44	61.66 +/- 10.30	**46.13 +/- 10.66**	**46.09 +/- 12.30**
ANOVA	70.95 +/- 3.38	55.21 +/- 4.44	**70.34 +/- 1.71**	61.14 +/- 19.79	40.08 +/- 15.43	32.60 +/- 16.84
15 Feature Set	68.59 +/- 3.63	54.23 +/- 5.28	66.05 +/- 11.60	**63.37 +/- 10.64**	42.59 +/- 10.51	42.32 +/- 11.98

Multiclass K-Nearest Neighbours

Feature Selection Method	AUC (%)	Balanced Accuracy (%)	Specificity (%)	Precision (%)	Sensitivity (%)	F1 Score (%)
MRMR	**73.28 +/- 3.46**	**59.30 +/- 5.72**	**68.82 +/- 7.09**	**67.06 +/- 5.47**	49.77 +/- 5.63	51.80 +/- 6.06
ANOVA	71.17 +/- 4.13	58.54 +/- 5.44	65.10 +/- 7.53	66.37 +/- 4.57	**51.98 +/- 4.53**	**54.52 +/- 4.49**
15 Feature Set	72.07 +/- 3.52	57.26 +/- 5.02	64.85 +/- 6.84	63.60 +/- 4.22	49.66 +/- 4.32	51.59 +/- 4.18

AUC is in a range from 68% to 78%, while the accuracy has a range from 54% to 62%. Only in the MLP model, the accuracy is above 60%. Between the feature selection methods, MRMR has slightly better results than the other two.

Table 5.2 presents the results of the multiclass experiments with STAD dataset. As in the previous table, MLP shows the best overall scores in all metrics. The AUC ranges from 68% to 81%, with values above 72% in just two of the three MLP experiments. The accuracy has a range from 51% to 63%, with the MLP having the best score with each feature selection method when compared to the Random Forest and the KNN. ANOVA method obtained the worst AUC and accuracy of all methods in all models.

Table 5.3 exhibits the results for the multiclass approach with the UCEC dataset. The MLP has the best AUC, always above 80%, while the accuracy is very similar between the MLP and the Random Forest. The AUC values are in a range from 76% to 85%, while the accuracy ones are in a range from 49% to 58%. Contrarily to the previous experiments, with the 15-feature set, MLP has worst results than Random Forest and KNN.

Table 5.2: Results of the STAD Dataset in the Multiclass approach.

STAD Dataset

Multiclass Multi-Layer Perceptron

Feature Selection Method	AUC (%)	Balanced Accuracy (%)	Specificity (%)	Precision (%)	Sensitivity (%)	F1 Score (%)
MRMR	**81.02 +/- 3.89**	**63.16 +/- 5.72**	**65.40 +/- 9.30**	**74.17 +/- 3.97**	**63.49 +/- 5.42**	**66.54 +/- 4.84**
ANOVA	72.07 +/- 5.49	55.99 +/- 6.95	53.09 +/- 9.91	66.17 +/- 4.96	55.98 +/- 6.76	58.96 +/- 6.29
15 Feature Set	75.81 +/- 4.56	58.01 +/- 6.32	63.29 +/- 10.25	68.94 +/- 5.26	53.78 +/- 8.63	59.92 +/- 8.08

Multiclass Random Forest

Feature Selection Method	AUC (%)	Balanced Accuracy (%)	Specificity (%)	Precision (%)	Sensitivity (%)	F1 Score (%)
MRMR	**72.35 +/- 4.79**	**56.68 +/- 5.25**	58.87 +/- 15.33	**71.77 +/- 6.01**	**52.49 +/- 10.82**	**55.75 +/- 10.41**
ANOVA	70.14 +/- 4.65	54.14 +/- 5.64	**59.67 +/- 11.58**	68.79 +/- 4.15	49.24 +/- 8.79	53.14 +/- 7.61
15 Feature Set	71.50 +/- 4.58	54.59 +/- 5.97	57.72 +/- 11.15	69.10 +/- 5.27	51.85 +/- 9.19	55.64 +/- 8.45

Multiclass K-Nearest Neighbours

Feature Selection Method	AUC (%)	Balanced Accuracy (%)	Specificity (%)	Precision (%)	Sensitivity (%)	F1 Score (%)
MRMR	**72.43 +/- 3.65**	56.68 +/- 4.85	59.11 +/- 6.46	**68.31 +/- 3.27**	**54.24 +/- 5.46**	**57.70 +/- 5.10**
ANOVA	68.96 +/- 4.99	51.70 +/- 6.95	52.85 +/- 9.88	65.19 +/- 4.46	50.56 +/- 5.48	53.88 +/- 5.10
15 Feature Set	71.81 +/- 4.91	**57.51 +/- 7.24**	**65.29 +/- 10.25**	66.82 +/- 5.86	49.73 +/- 5.88	53.06 +/- 5.76

Table 5.3: Results of the UCEC Dataset in the Multiclass approach.

UCEC Dataset

Multiclass Multi-Layer Perceptron

Feature Selection Method	AUC (%)	Balanced Accuracy (%)	Specificity (%)	Precision (%)	Sensitivity (%)	F1 Score (%)
MRMR	**85.28 +/- 2.63**	**58.48 +/- 5.34**	50.55 +/- 8.33	**77.31 +/- 3.45**	**63.64 +/- 6.06**	**68.31 +/- 5.40**
ANOVA	82.71 +/- 2.80	56.79 +/- 5.55	50.00 +/- 9.61	74.81 +/- 3.21	63.58 +/- 3.81	67.62 +/- 3.20
15 Feature Set	80.24 +/- 3.32	54.20 +/- 5.03	**53.92 +/- 8.03**	72.28 +/- 4.08	54.48 +/- 6.51	58.08 +/- 7.26

Multiclass Random Forest

Feature Selection Method	AUC (%)	Balanced Accuracy (%)	Specificity (%)	Precision (%)	Sensitivity (%)	F1 Score (%)
MRMR	**79.48 +/- 4.00**	**57.74 +/- 6.39**	53.66 +/- 11.30	**77.12 +/- 4.60**	**61.81 +/- 7.57**	**66.68 +/- 6.09**
ANOVA	79.44 +/- 3.95	55.98 +/- 4.90	50.97 +/- 11.43	75.52 +/- 4.52	60.99 +/- 8.45	65.18 +/- 7.12
15 Feature Set	76.68 +/- 3.70	56.89 +/- 5.42	**59.43 +/- 11.69**	74.49 +/- 4.70	54.34 +/- 8.77	59.16 +/- 8.71

Multiclass K-Nearest Neighbours

Feature Selection Method	AUC (%)	Balanced Accuracy (%)	Specificity (%)	Precision (%)	Sensitivity (%)	F1 Score (%)
MRMR	79.54 +/- 3.41	54.26 +/- 6.12	50.21 +/- 9.66	72.30 +/- 3.69	58.32 +/- 4.90	62.69 +/- 4.44
ANOVA	77.21 +/- 2.92	49.59 +/- 7.16	49.13 +/- 11.09	71.16 +/- 4.24	50.06 +/- 5.15	55.42 +/- 4.96
15 Feature Set	**81.13 +/- 3.59**	**56.38 +/- 6.01**	**53.48 +/- 10.04**	**75.24 +/- 3.90**	**59.28 +/- 4.54**	**64.25 +/- 4.18**

5.1.2 Binary Approach: Merging MSI-L with MSI-H

Table 5.4 presents the results of the binary classification experiments with the COAD dataset, in which MSI-L class is merged with MSI-H class. MLP has the best results overall, and Random Forest has the worst. The AUC ranges from 69% to 79%, while the accuracy ranges from 64% to 72%. The method that combines the Dimensionality Reduction (DR) technique with ANOVA has the best AUC and accuracy in the Random Forest and the KNN models. However, its values are below DR + MRMR and MRMR in the MLP.

Table 5.5 exhibits the results of the experiments with the STAD dataset in the binary classification model in which MSI-L class is merged with MSI-H class. MLP has the best results overall, except for the DR + ANOVA method, which is the worst of the three models. The AUC ranges from 71% to 81%, while the accuracy ranges from 65% to 74%. The method that combines DR with MRMR has the best results in MLP and KNN, while in the Random Forest is the method with DR and ANOVA that produces the best results.

Table 5.6 shows the results for the binary classification model that merges the MSI-L and MSI-H labels for the UCEC dataset. The MLP model obtained the best results with four feature selection methods, with the only exception being the 15-feature set that produced the best result with the KNN model. The AUC ranges from 80% to 88%, while the accuracy ranges from 75% to 82%, with the DR + MRMR and the MRMR methods in the MLP model obtaining the best performance of the set of experiments.

5.1.3 Binary Approach: Merging MSI-L with MSS

Table 5.7 presents the results of the experiments in the binary classification approach that merges MSI-L with MSS using the COAD dataset. The AUC values range from 89% to 98%, and the accuracy values range from 85% to 92%. The approach with DR + MRMR feature selection method in the MLP obtained the best results in all metrics. The approach with the ANOVA method was the best in the Random Forest experiments, but their results are below any experiment in the

Table 5.4: Results of the COAD dataset in the binary classification approach that merges MSI-L with MSI-H.

COAD Dataset

Binary Multi-Layer Perceptron (merge of MSI-L with MSI-H)

Feature Selection Method	**AUC (%)**	**Balanced Accuracy (%)**	**Specificity (%)**	**Precision (%)**	**Sensitivity (%)**	**F1 Score (%)**
DR + MRMR	78.50 +/- 3.85	**72.30 +/- 4.00**	68.79 +/- 4.50	74.05 +/- 3.54	74.00 +/- 3.63	73.71 +/- 3.57
DR + ANOVA	78.16 +/- 4.52	71.65 +/- 3.88	**69.76 +/- 4.53**	**74.66 +/- 3.59**	74.57 +/- 3.53	**74.34 +/- 3.49**
MRMR	**79.24 +/- 4.30**	71.79 +/- 3.99	69.19 +/- 4.99	73.07 +/- 3.92	72.69 +/- 3.84	72.70 +/- 3.82
ANOVA	76.69 +/- 4.66	71.45 +/- 4.24	68.13 +/- 4.85	74.54 +/- 4.11	**74.76 +/- 3.98**	74.24 +/- 4.01
15 Feature Set	73.79 +/- 4.82	69.08 +/- 4.65	66.39 +/- 5.35	71.86 +/- 4.42	71.77 +/- 4.56	71.46 +/- 4.46

Binary Random Forest (merge of MSI-L with MSI-H)

Feature Selection Method	**AUC (%)**	**Balanced Accuracy (%)**	**Specificity (%)**	**Precision (%)**	**Sensitivity (%)**	**F1 Score (%)**
DR + MRMR	70.14 +/- 5.17	67.45 +/- 4.06	63.43 +/- 5.30	69.53 +/- 5.52	68.78 +/- 6.40	68.36 +/- 5.95
DR + ANOVA	**72.52 +/- 4.69**	**68.60 +/- 4.43**	64.48 +/- 4.00	71.71 +/- 5.30	70.46 +/- 5.93	69.81 +/- 5.22
MRMR	72.41 +/- 4.85	67.85 +/- 4.28	**65.41 +/- 5.16**	71.47 +/- 4.13	70.29 +/- 5.50	69.75 +/- 5.16
ANOVA	69.57 +/- 5.17	66.53 +/- 4.68	63.41 +/- 4.53	69.96 +/- 5.06	69.64 +/- 5.71	69.15 +/- 5.25
15 Feature Set	71.41 +/- 6.05	67.93 +/- 5.19	63.39 +/- 5.49	**73.04 +/- 6.01**	**72.46 +/- 5.88**	**71.27 +/- 5.50**

Binary K-Nearest Neighbours (merge of MSI-L with MSI-H)

Feature Selection Method	**AUC (%)**	**Balanced Accuracy (%)**	**Specificity (%)**	**Precision (%)**	**Sensitivity (%)**	**F1 Score (%)**
DR + MRMR	75.09 +/- 4.79	67.50 +/- 4.29	68.72 +/- 4.83	**70.85 +/- 4.17**	**68.95 +/- 4.63**	**69.34 +/- 4.52**
DR + ANOVA	**76.21 +/- 4.21**	**70.24 +/- 4.01**	**69.27 +/- 4.31**	70.73 +/- 3.63	67.89 +/- 3.80	68.38 +/- 3.71
MRMR	72.11 +/- 4.76	65.61 +/- 4.37	65.88 +/- 4.83	68.02 +/- 4.08	65.35 +/- 4.45	65.81 +/- 4.33
ANOVA	75.47 +/- 4.43	68.68 +/- 4.01	69.20 +/- 4.25	70.75 +/- 3.68	68.15 +/- 4.07	68.63 +/- 3.99
15 Feature Set	70.47 +/- 4.72	64.93 +/- 4.94	63.42 +/- 5.34	67.41 +/- 4.63	66.44 +/- 5.06	66.60 +/- 4.89

Table 5.5: Results of the STAD dataset in the binary classification approach that merges MSI-L with MSI-H.

STAD Dataset

Binary Multi-Layer Perceptron (merge of MSI-L with MSI-H)

Feature Selection Method	**AUC (%)**	**Balanced Accuracy (%)**	**Specificity (%)**	**Precision (%)**	**Sensitivity (%)**	**F1 Score (%)**
DR + MRMR	**81.29 +/- 4.76**	**74.09 +/- 4.60**	**72.53 +/- 5.80**	**77.25 +/- 4.26**	**76.15 +/- 4.72**	**76.35 +/- 4.52**
DR + ANOVA	73.39 +/- 5.40	69.25 +/- 3.76	65.39 +/- 6.32	70.64 +/- 4.46	68.59 +/- 4.85	69.08 +/- 4.67
MRMR	80.08 +/- 5.25	74.01 +/- 5.23	71.35 +/- 6.14	76.48 +/- 4.49	75.51 +/- 4.85	75.69 +/- 4.68
ANOVA	74.14 +/- 5.44	68.79 +/- 4.49	64.85 +/- 5.55	70.87 +/- 4.01	69.49 +/- 4.39	69.82 +/- 4.19
15 Feature Set	74.88 +/- 5.62	69.75 +/- 4.73	65.87 +/- 6.01	72.71 +/- 4.23	72.02 +/- 4.37	72.07 +/- 4.20

Binary Random Forest (merge of MSI-L with MSI-H)

Feature Selection Method	**AUC (%)**	**Balanced Accuracy (%)**	**Specificity (%)**	**Precision (%)**	**Sensitivity (%)**	**F1 Score (%)**
DR + MRMR	74.97 +/- 5.48	**70.33 +/- 5.18**	67.77 +/- 6.05	73.41 +/- 5.18	70.90 +/- 6.81	71.15 +/- 6.27
DR + ANOVA	**75.31 +/- 5.72**	70.02 +/- 5.48	**67.82 +/- 6.31**	**74.99 +/- 5.01**	**73.85 +/- 6.13**	**73.76 +/- 5.68**
MRMR	74.09 +/- 5.52	68.99 +/- 5.69	67.56 +/- 6.38	73.08 +/- 5.03	70.41 +/- 7.10	70.65 +/- 6.62
ANOVA	72.08 +/- 5.50	67.10 +/- 4.49	61.22 +/- 6.61	72.58 +/- 5.16	70.78 +/- 6.92	69.81 +/- 6.03
15 Feature Set	72.17 +/- 4.59	68.19 +/- 4.61	65.26 +/- 5.57	73.17 +/- 5.04	71.41 +/- 6.44	71.27 +/- 5.83

Binary K-Nearest Neighbours (merge of MSI-L with MSI-H)

Feature Selection Method	**AUC (%)**	**Balanced Accuracy (%)**	**Specificity (%)**	**Precision (%)**	**Sensitivity (%)**	**F1 Score (%)**
DR + MRMR	**76.87 +/- 5.57**	**70.26 +/- 5.18**	70.06 +/- 5.55	**72.62 +/- 3.96**	**68.12 +/- 4.36**	**68.92 +/- 4.19**
DR + ANOVA	75.20 +/- 5.33	69.03 +/- 5.88	**70.40 +/- 5.51**	72.43 +/- 4.22	66.83 +/- 4.97	67.69 +/- 4.82
MRMR	73.43 +/- 4.66	66.56 +/- 4.62	65.32 +/- 5.27	70.24 +/- 3.88	67.76 +/- 4.46	68.36 +/- 4.20
ANOVA	71.02 +/- 4.85	65.32 +/- 5.00	63.88 +/- 6.04	69.09 +/- 4.33	66.76 +/- 4.50	67.40 +/- 4.35
15 Feature Set	72.46 +/- 4.91	65.61 +/- 4.88	64.07 +/- 5.77	69.47 +/- 4.21	67.22 +/- 4.84	67.77 +/- 4.63

Table 5.6: Results of the UCEC dataset in the binary classification approach that merges MSI-L with MSI-H.

UCEC Dataset

Binary Multi-Layer Perceptron (merge of MSI-L with MSI-H)

Feature Selection Method	AUC (%)	Balanced Accuracy (%)	Specificity (%)	Precision (%)	Sensitivity (%)	F1 Score (%)
DR + MRMR	**88.57 +/- 3.55**	82.79 +/- 3.90	**81.36 +/- 4.34**	82.76 +/- 3.99	82.55 +/- 3.98	82.52 +/- 3.97
DR + ANOVA	87.69 +/- 3.29	82.21 +/- 3.47	80.04 +/- 3.89	81.82 +/- 3.65	81.66 +/- 3.67	81.59 +/- 3.65
MRMR	88.50 +/- 3.50	**82.92 +/- 3.88**	80.95 +/- 3.67	**83.36 +/- 2.92**	**83.14 +/- 2.92**	**82.99 +/- 2.98**
ANOVA	87.61 +/- 3.50	81.55 +/- 3.40	80.71 +/- 3.68	82.06 +/- 3.23	81.81 +/- 3.33	81.79 +/- 3.34
15 Feature Set	81.49 +/- 4.08	76.07 +/- 4.23	73.95 +/- 4.76	78.47 +/- 4.64	78.00 +/- 4.30	77.52 +/- 4.34

Binary Random Forest (merge of MSI-L with MSI-H)

Feature Selection Method	AUC (%)	Balanced Accuracy (%)	Specificity (%)	Precision (%)	Sensitivity (%)	F1 Score (%)
DR + MRMR	**83.20 +/- 4.10**	76.25 +/- 4.79	76.18 +/- 4.38	78.61 +/- 4.13	77.73 +/- 4.71	77.55 +/- 4.72
DR + ANOVA	82.64 +/- 4.17	77.00 +/- 3.79	76.01 +/- 4.73	78.49 +/- 4.10	77.89 +/- 3.99	77.71 +/- 3.99
MRMR	82.76 +/- 4.33	76.86 +/- 4.28	75.68 +/- 4.47	78.86 +/- 4.43	77.96 +/- 4.54	77.70 +/- 4.44
ANOVA	82.22 +/- 5.04	**77.27 +/- 4.45**	**76.25 +/- 4.92**	78.79 +/- 4.23	78.42 +/- 4.13	78.24 +/- 4.17
15 Feature Set	80.25 +/- 4.51	76.44 +/- 4.19	74.23 +/- 4.53	**79.45 +/- 4.49**	**78.65 +/- 4.21**	78.07 +/- 4.21

Binary K-Nearest Neighbours (merge of MSI-L with MSI-H)

Feature Selection Method	AUC (%)	Balanced Accuracy (%)	Specificity (%)	Precision (%)	Sensitivity (%)	F1 Score (%)
DR + MRMR	84.61 +/- 3.80	75.94 +/- 4.56	77.43 +/- 4.78	78.34 +/- 4.12	77.77 +/- 4.04	77.83 +/- 4.06
DR + ANOVA	**85.28 +/- 4.37**	76.52 +/- 5.00	**78.03 +/- 4.97**	**79.61 +/- 4.30**	**79.28 +/- 4.24**	**79.23 +/- 4.25**
MRMR	85.02 +/- 4.54	**77.79 +/- 4.37**	77.65 +/- 4.98	78.52 +/- 4.12	77.92 +/- 3.89	77.97 +/- 3.94
ANOVA	83.81 +/- 4.48	77.19 +/- 4.54	76.39 +/- 5.00	78.11 +/- 4.25	78.00 +/- 4.17	77.94 +/- 4.24
15 Feature Set	83.99 +/- 3.97	77.59 +/- 3.78	76.18 +/- 4.20	78.95 +/- 3.60	78.90 +/- 3.59	78.68 +/- 3.68

MLP model. The MRMR method stood out in the KNN model, having the best results of all methods in that model.

Table 5.8 exhibits the results of the binary classification experiments with the STAD dataset, in which MSI-L class is merged with the MSS class. The AUC is in a range from 79% to 95%, and the accuracy is in a range from 78% to 89%. The 15-feature set obtained the best results in the MLP and the KNN models in the two best experiments of this approach. In the Random Forest, it was the ANOVA method that performed better. With the ANOVA and the DR + ANOVA methods, the KNN obtained the worst results from this approach.

Table 5.9 shows the results of the experiments with the UCEC dataset in the binary classification model in which MSI-L class is merged with the MSS class. The AUC has a range of values from 84% to 94%, while the accuracy has a range from 76% to 88%. The best results are obtained with the DR + MRMR and the MRMR methods in the MLP model, with both experiments reaching an AUC of 94% and an accuracy of 87% and 88%, respectively. Random Forest is the model with the worst overall AUC values, and KNN is the one with the worst overall accuracy. However, in both of these models, the experiment with the 15-feature set obtained a similar result to the experiment with that same method in the MLP model.

5.2 Genetic Selection Results Analysis

Five methods to select features were applied, two of which were composed of two techniques. The three types of cancer studied, and the three different approaches would result in 75 different experiments in feature selection. Instead, it only represents 63 experiments since the 15-feature set from [19] always selects the same group of genes from the three different datasets. The analysis

Table 5.7: Results of the COAD dataset in the binary classification approach that merges MSI-L with MSS.

COAD Dataset						
Binary Multi-Layer Perceptron (merge of MSI-L with MSS)						
Feature Selection Method	**AUC (%)**	**Balanced Accuracy (%)**	**Specificity (%)**	**Precision (%)**	**Sensitivity (%)**	**F1 Score (%)**
DR + MRMR	**98.44 +/- 1.20**	**92.67 +/- 3.07**	**93.91 +/- 4.27**	**94.40 +/- 1.82**	93.24 +/- 2.42	**93.52 +/- 2.25**
DR + ANOVA	97.53 +/- 1.61	92.49 +/- 3.56	92.52 +/- 5.19	93.73 +/- 2.12	92.44 +/- 2.77	92.76 +/- 2.57
MRMR	97.62 +/- 1.56	92.28 +/- 3.92	89.33 +/- 6.01	93.85 +/- 2.49	**93.31 +/- 2.78**	93.45 +/- 2.67
ANOVA	97.12 +/- 1.54	91.25 +/- 3.25	91.03 +/- 4.79	93.00 +/- 2.04	91.47 +/- 2.89	91.86 +/- 2.64
15 Feature Set	96.19 +/- 2.00	89.44 +/- 3.70	89.03 +/- 5.96	91.74 +/- 1.97	89.85 +/- 2.56	90.34 +/- 2.33

Binary Random Forest (merge of MSI-L with MSS)						
Feature Selection Method	**AUC (%)**	**Balanced Accuracy (%)**	**Specificity (%)**	**Precision (%)**	**Sensitivity (%)**	**F1 Score (%)**
DR + MRMR	90.53 +/- 5.32	87.78 +/- 5.02	83.92 +/- 7.58	91.22 +/- 2.60	90.21 +/- 3.15	90.47 +/- 2.92
DR + ANOVA	89.78 +/- 5.33	87.13 +/- 3.89	83.07 +/- 7.13	90.58 +/- 2.95	89.49 +/- 3.42	89.82 +/- 3.20
MRMR	**90.97 +/- 4.60**	86.86 +/- 4.67	84.55 +/- 6.84	90.52 +/- 2.74	88.55 +/- 4.23	89.07 +/- 3.76
ANOVA	90.83 +/- 4.52	**87.86 +/- 4.71**	**85.25 +/- 7.36**	**91.53 +/- 2.66**	**90.46 +/- 3.15**	**90.75 +/- 2.94**
15 Feature Set	**90.97 +/- 4.60**	85.17 +/- 3.45	84.55 +/- 6.84	90.52 +/- 2.74	88.55 +/- 4.23	89.07 +/- 4.60

Binary K-Nearest Neighbours (merge of MSI-L with MSS)						
Feature Selection Method	**AUC (%)**	**Balanced Accuracy (%)**	**Specificity (%)**	**Precision (%)**	**Sensitivity (%)**	**F1 Score (%)**
DR + MRMR	95.11 +/- 2.51	87.85 +/- 4.24	88.28 +/- 7.25	90.67 +/- 2.39	87.73 +/- 3.27	88.46 +/- 2.94
DR + ANOVA	95.32 +/- 2.30	89.91 +/- 3.79	88.38 +/- 6.58	89.99 +/- 1.98	86.00 +/- 2.88	86.98 +/- 2.53
MRMR	**95.77 +/- 2.02**	**91.63 +/- 2.89**	**93.03 +/- 4.01**	**92.73 +/- 1.88**	**90.23 +/- 3.01**	**90.82 +/- 2.70**
ANOVA	93.48 +/- 3.20	88.70 +/- 4.08	87.92 +/- 6.13	91.32 +/- 2.31	89.47 +/- 2.84	89.98 +/- 2.62
15 Feature Set	93.91 +/- 2.64	88.47 +/- 3.47	88.80 +/- 5.20	90.94 +/- 2.03	88.15 +/- 3.21	88.87 +/- 2.84

Table 5.8: Results of the STAD dataset in the binary classification approach that merges MSI-L with MSS.

STAD Dataset						
Binary Multi-Layer Perceptron (merge of MSI-L with MSS)						
Feature Selection Method	**AUC (%)**	**Balanced Accuracy (%)**	**Specificity (%)**	**Precision (%)**	**Sensitivity (%)**	**F1 Score (%)**
DR + MRMR	91.80 +/- 5.63	87.00 +/- 5.01	81.17 +/- 9.52	91.21 +/- 3.13	90.76 +/- 3.19	90.85 +/- 3.15
DR + ANOVA	89.82 +/- 5.05	85.65 +/- 5.51	78.28 +/- 9.34	90.06 +/- 3.35	89.44 +/- 3.81	89.58 +/- 3.61
MRMR	93.06 +/- 4.41	87.88 +/- 4.70	83.61 +/- 7.45	**92.37 +/- 2.53**	**91.98 +/- 2.64**	**92.06 +/- 2.56**
ANOVA	88.77 +/- 5.93	82.00 +/- 5.48	76.94 +/- 9.47	88.34 +/- 2.92	87.12 +/- 3.00	87.49 +/- 2.87
15 Feature Set	**95.78 +/- 2.07**	**89.17 +/- 3.93**	**86.81 +/- 6.20**	92.10 +/- 2.22	91.00 +/- 2.67	91.32 +/- 2.49

Binary Random Forest (merge of MSI-L with MSS)						
Feature Selection Method	**AUC (%)**	**Balanced Accuracy (%)**	**Specificity (%)**	**Precision (%)**	**Sensitivity (%)**	**F1 Score (%)**
DR + MRMR	87.08 +/- 7.09	85.66 +/- 5.41	78.80 +/- 1.24	90.25 +/- 2.71	88.98 +/- 2.91	89.12 +/- 2.71
DR + ANOVA	**87.98 +/- 7.87**	85.42 +/- 6.18	81.30 +/- 10.30	91.17 +/- 3.03	90.44 +/- 3.15	90.57 +/- 3.08
MRMR	86.99 +/- 6.99	85.41 +/- 5.83	79.90 +/- 1.08	91.16 +/- 2.82	80.63 +/- 2.74	80.65 +/- 2.78
ANOVA	87.78 +/- 7.23	**86.97 +/- 5.81**	**82.42 +/- 10.28**	**91.88 +/- 2.53**	**91.27 +/- 2.52**	**91.34 +/- 2.52**
15 Feature Set	86.70 +/- 4.89	82.25 +/- 4.88	77.44 +/- 8.81	88.89 +/- 2.65	87.54 +/- 3.25	87.89 +/- 2.93

Binary K-Nearest Neighbours (merge of MSI-L with MSS)						
Feature Selection Method	**AUC (%)**	**Balanced Accuracy (%)**	**Specificity (%)**	**Precision (%)**	**Sensitivity (%)**	**F1 Score (%)**
DR + MRMR	86.59 +/- 4.98	83.15 +/- 5.10	79.91 +/- 8.42	85.62 +/- 2.67	77.76 +/- 4.11	79.81 +/- 3.49
DR + ANOVA	79.11 +/- 5.78	79.27 +/- 5.83	71.02 +/- 10.29	82.33 +/- 3.31	74.54 +/- 3.89	76.83 +/- 3.36
MRMR	92.56 +/- 4.60	87.33 +/- 4.32	88.35 +/- 7.30	90.24 +/- 2.23	86.32 +/- 2.99	87.30 +/- 2.66
ANOVA	85.44 +/- 6.36	78.30 +/- 5.75	78.53 +/- 9.00	85.34 +/- 3.10	78.07 +/- 4.87	80.04 +/- 4.18
15 Feature Set	**94.14 +/- 2.79**	**89.75 +/- 3.38**	**89.23 +/- 5.37**	**92.16 +/- 1.97**	**90.27 +/- 2.93**	**90.77 +/- 2.62**

Table 5.9: Results of the UCEC dataset in the binary classification approach that merges MSI-L with MSS.

UCEC Dataset

Binary Multi-Layer Perceptron (merge of MSI-L with MSS)

Feature Selection Method	AUC (%)	Balanced Accuracy (%)	Specificity (%)	Precision (%)	Sensitivity (%)	F1 Score (%)
DR + MRMR	94.41 +/- 2.38	87.50 +/- 3.44	86.17 +/- 4.44	89.09 +/- 2.95	88.72 +/- 3.18	88.77 +/- 3.12
DR + ANOVA	92.86 +/- 2.25	85.54 +/- 3.40	83.02 +/- 4.14	86.23 +/- 2.82	85.52 +/- 3.18	85.65 +/- 3.06
MRMR	**94.47 +/- 2.22**	**88.07 +/- 3.13**	**86.66 +/- 4.37**	**89.23 +/- 2.72**	**88.72 +/- 2.98**	**88.79 +/- 2.92**
ANOVA	91.87 +/- 2.74	84.40 +/- 3.27	82.72 +/- 4.12	86.26 +/- 2.70	85.64 +/- 3.04	85.73 +/- 2.93
15 Feature Set	87.86 +/- 4.20	82.13 +/- 3.55	80.63 +/- 5.30	83.97 +/- 3.44	82.95 +/- 3.67	83.15 +/- 3.56

Binary Random Forest (merge of MSI-L with MSS)

Feature Selection Method	AUC (%)	Balanced Accuracy (%)	Specificity (%)	Precision (%)	Sensitivity (%)	F1 Score (%)
DR + MRMR	85.86 +/- 4.15	80.07 +/- 4.27	78.76 +/- 5.59	82.30 +/- 3.85	82.38 +/- 4.29	82.49 +/- 4.14
DR + ANOVA	85.39 +/- 3.79	80.26 +/- 3.96	78.30 +/- 4.37	82.78 +/- 3.22	81.81 +/- 3.76	81.96 +/- 3.53
MRMR	84.80 +/- 4.84	80.79 +/- 4.57	78.22 +/- 5.45	82.76 +/- 3.71	81.92 +/- 4.10	82.05 +/- 3.95
ANOVA	86.10 +/- 4.44	80.84 +/- 4.62	78.47 +/- 5.48	84.10 +/- 4.38	83.31 +/- 4.68	83.32 +/- 4.45
15 Feature Set	**86.53 +/- 3.99**	**82.33 +/- 3.62**	**79.66 +/- 5.00**	**85.30 +/- 2.93**	**84.99 +/- 3.07**	**84.91 +/- 3.04**

Binary K-Nearest Neighbours (merge of MSI-L with MSS)

Feature Selection Method	AUC (%)	Balanced Accuracy (%)	Specificity (%)	Precision (%)	Sensitivity (%)	F1 Score (%)
DR + MRMR	87.82 +/- 3.47	80.18 +/- 4.12	80.01 +/- 4.70	80.84 +/- 3.28	77.01 +/- 3.78	77.72 +/- 3.61
DR + ANOVA	87.25 +/- 3.38	78.04 +/- 3.49	78.27 +/- 5.20	79.95 +/- 3.32	76.59 +/- 3.46	77.25 +/- 3.33
MRMR	88.49 +/- 3.19	79.53 +/- 4.45	79.76 +/- 5.42	81.10 +/- 3.75	78.11 +/- 4.09	78.73 +/- 3.95
ANOVA	84.91 +/- 3.55	76.93 +/- 3.93	78.48 +/- 4.57	79.55 +/- 3.26	75.39 +/- 3.82	76.17 +/- 3.66
15 Feature Set	**89.57 +/- 3.63**	**82.73 +/- 3.94**	**81.06 +/- 5.54**	**84.99 +/- 3.07**	**84.40 +/- 2.93**	**84.47 +/- 2.94**

of the MRMR method is made in subsection 5.2.1, while subsection 5.2.2 contains the analysis of the ANOVA method. Subsections 5.2.3 and 5.2.4 analyse these two methods combined with the dimensionality reduction technique, while subsection 5.2.5 analyses the selection of the 15-feature set.

The feature selection analysis will be performed with the multiclass approach and the binary classification with the merge of MSI-L with MSS since it obtained better results when compared to the other binary classification approach. From the results in the previous section, eight different experiments were highlighted to analyse the feature selection results. Table 5.10 details the eight approaches selected. All the others are in Appendix C.

The choice of experiments for analysis was based on the best results of each method in the two approaches, but also guaranteeing that each dataset was selected at least once in each approach. Thus, the selection will have one analysis for MRMR, ANOVA and the 15-feature set, all with distinct datasets, and one analysis for the five methods, generating the eight selected experiments.

Table 5.10: Selected experiments for feature selection analysis.

Experiment	Dataset	Approach	Feature Selection Method
1	STAD	Multiclass	MRMR
2	UCEC	Binary (merging MSI-L with MSS)	MRMR
3	COAD	Multiclass	ANOVA
4	COAD	Binary (merging MSI-L with MSS)	ANOVA
5	COAD	Binary (merging MSI-L with MSS)	DR + MRMR
6	STAD	Binary (merging MSI-L with MSS)	DR + ANOVA
7	UCEC	Multiclass	15 Feature Set
8	UCEC	Binary (merging MSI-L with MSS)	15 Feature Set

5.2.1 MRMR

Figure 5.1 presents the heatmap containing the expression of the genes with most variation from the 100 selected by MRMR in the multiclass approach with the STAD dataset. The heatmap shows genes *SERPINB5*, *RPL22L1*, *CXCL1* and *AFAP1-AS1* are more expressed in the MSI-H samples, while genes *ATP5F1A*, *ATP5MC3*, *PRDX2* and *TSPO* are more expressed in the MSI-L and MSS samples. On the other hand, *CDC42EP1* is highly expressed in all patients.

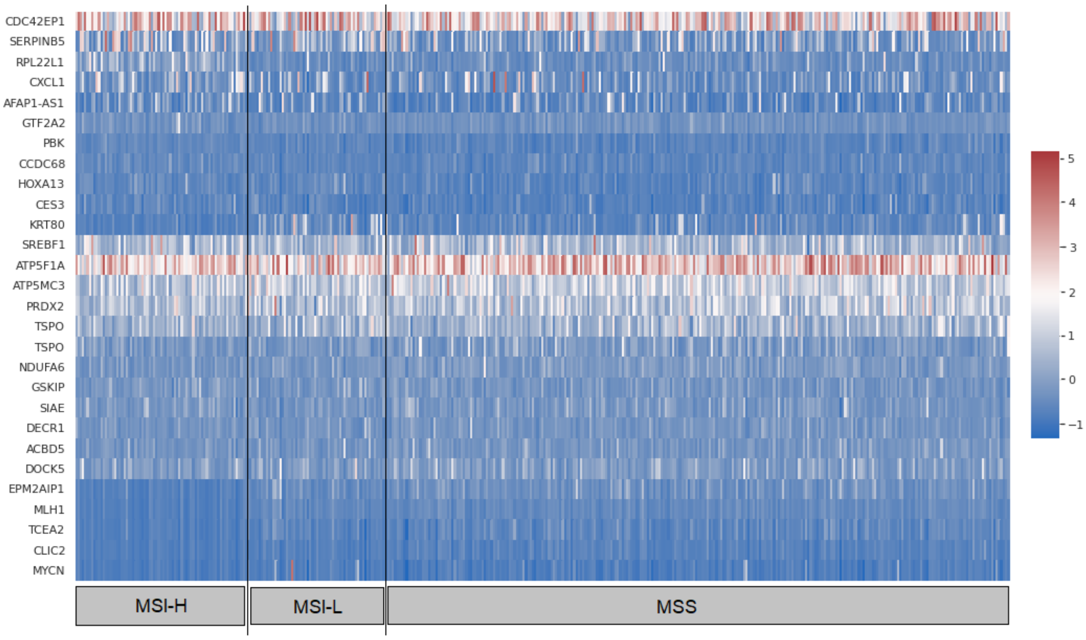

Figure 5.1: Heatmap of the most influential genes selected by MRMR with the 3class STAD dataset.

The heatmap with the expression of most relevant genes from the 100 selected by MRMR in the binary classification approach with the UCEC dataset is exhibited in figure 5.2. Genes *RBP1*, *MYO1C*, *LAD1*, *CDKN2A* and *ERO1A* are more expressed in patients with MSS or MSI-L, while genes *MSX1* and *ANXA1* are strongly associated with patients in MSI-h condition.

This method's runtime was 16 minutes with COAD and STAD, and 17 minutes with UCEC in the multiclass approach. This slight variation explains itself by the number of samples in each dataset. In the binary classification approaches, the runtime was 19 minutes with COAD and STAD, and 20 minutes with UCEC.

5.2.2 ANOVA

Figure 5.3 shows the heatmap expressing the most important genes from the 100 selected in the multiclass approach with the COAD dataset. Gene *AGR2* stands out as much more expressed in MSI-H samples when compared with MSI-L and MSS samples, while genes *RNF43*, *TMEM176B* and *CFTR* denote themselves in the MSI-L and MSS samples. *ATP5F1A*, *WARS1* and *CD55* also denothe themselves in MSI-H patients.

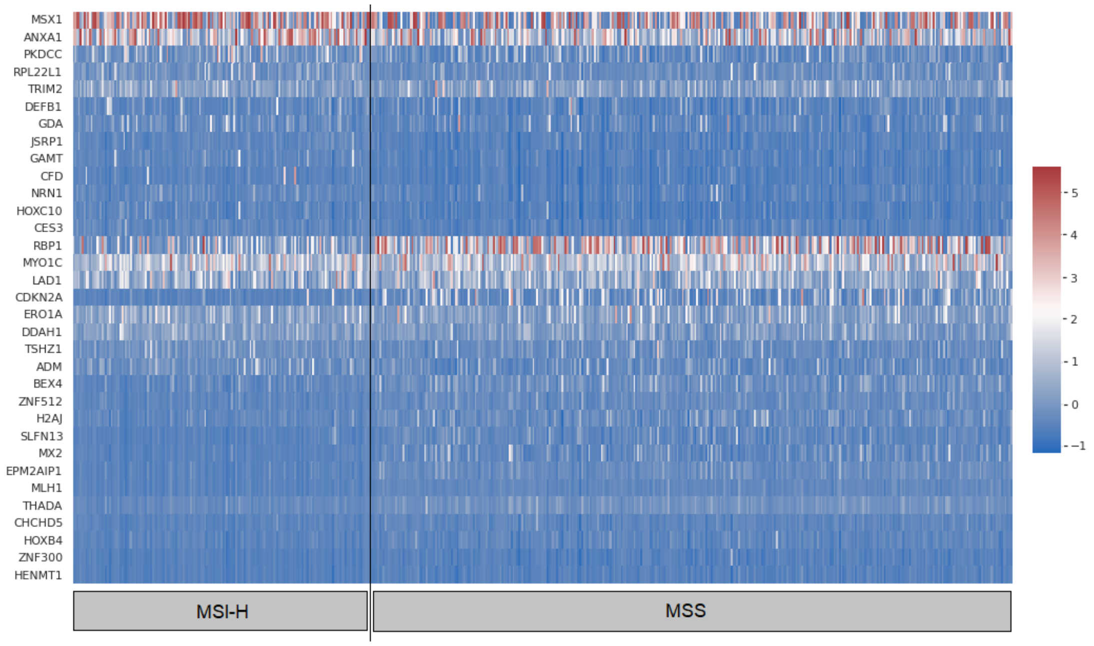

Figure 5.2: Heatmap of the most influential genes selected by MRMR with the 2class UCEC dataset.

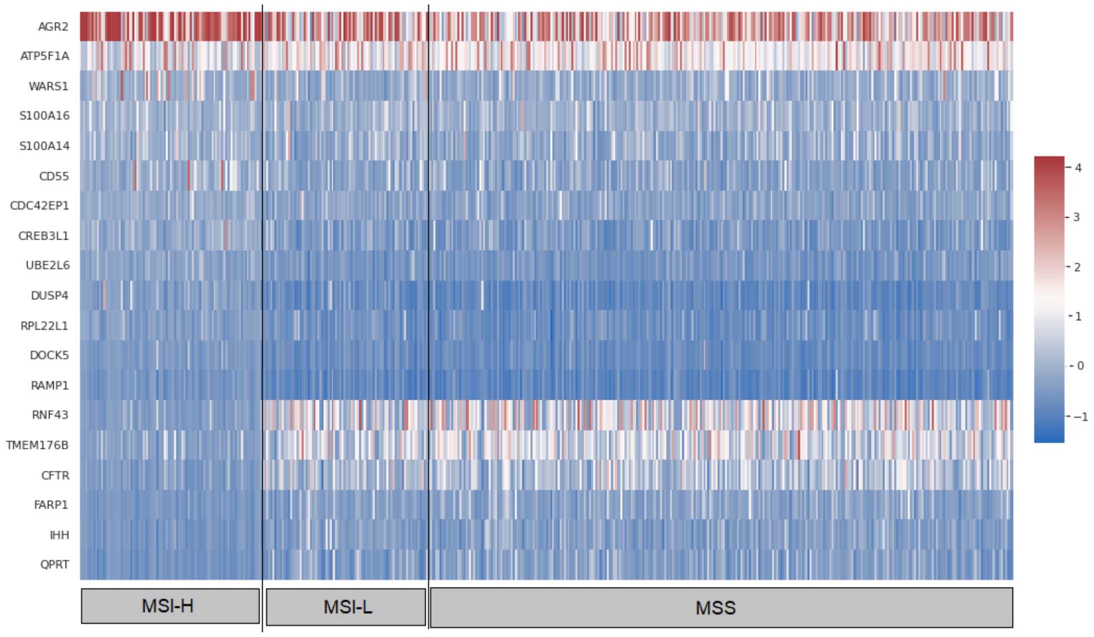

Figure 5.3: Heatmap of the most influential genes selected by ANOVA with the 3class COAD dataset.

The heatmap with the expression levels of the genes with the bigger ranges from the 100 selected in the binary classification approach with the COAD dataset is presented in figure 5.4. Gene *AGR2* highlights itself in MSI-H samples, while genes *QRPT*, *RNF43* and *TMEM176B* have more representation in MSS samples. In smaller variations, *CD55* and *DUSP4* are more expressed in MSI-H patients.

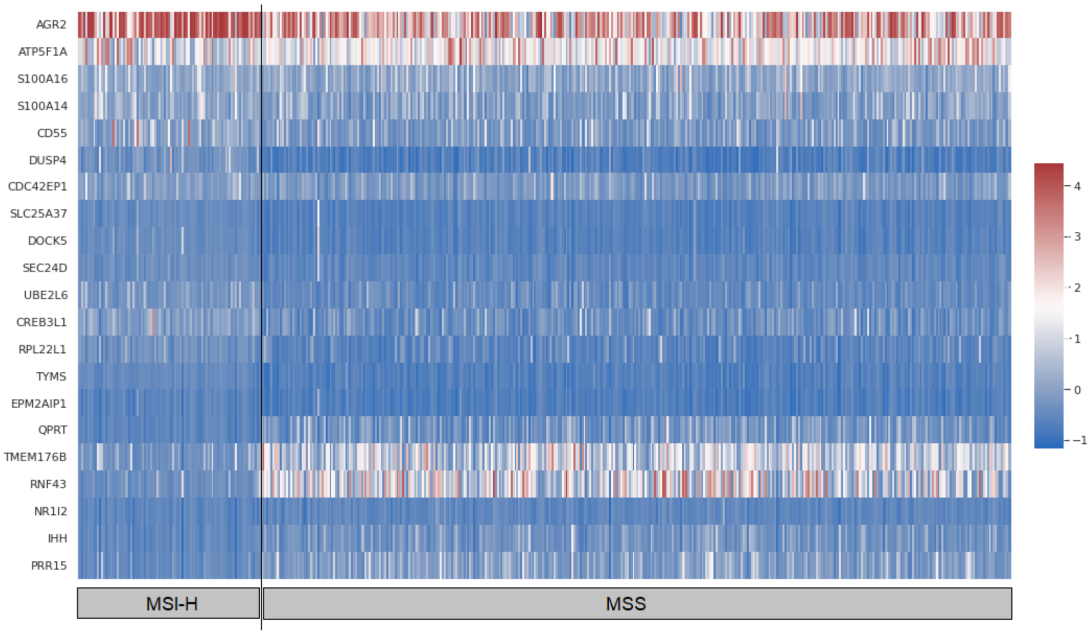

Figure 5.4: Heatmap of the most influential genes selected by ANOVA with the 2class COAD dataset.

This method's runtime was 1 second with COAD and STAD, and 2 seconds with UCEC in all approaches.

5.2.3 Dimensionality Reduction and MRMR

Figure 5.5 represents one of the most important images of this section, with the heatmap of the features that produced the approach with the better result overall in a combination of DR and MRMR with the COAD dataset. Genes *S100P*, *CREB3L1*, *S100A16*, *CDC42EP1*, *CXCL16*, *UBE2L6*, *PFKP*, *RPL22L1*, *DUSP4*, *SREBF1* and *GBP2* stand out in the MSI-H patients, while MSS patients have a big group of genes that have more expression. It includes *RNF43*, *TMEM176B*, *ADGRG1*, *CHMP4B*, *ATP9A*, *TMEM176A*, *CFTR*, *POFUT1*, *NOX1* and *TSPAN6*.

This method's runtime was 18 minutes with COAD and UCEC, and 19 minutes with STAD. The DR method needed between 15 to 16 minutes to run, while the MRMR was much faster, running in 3 to 4 minutes since it had fewer features to choose from.

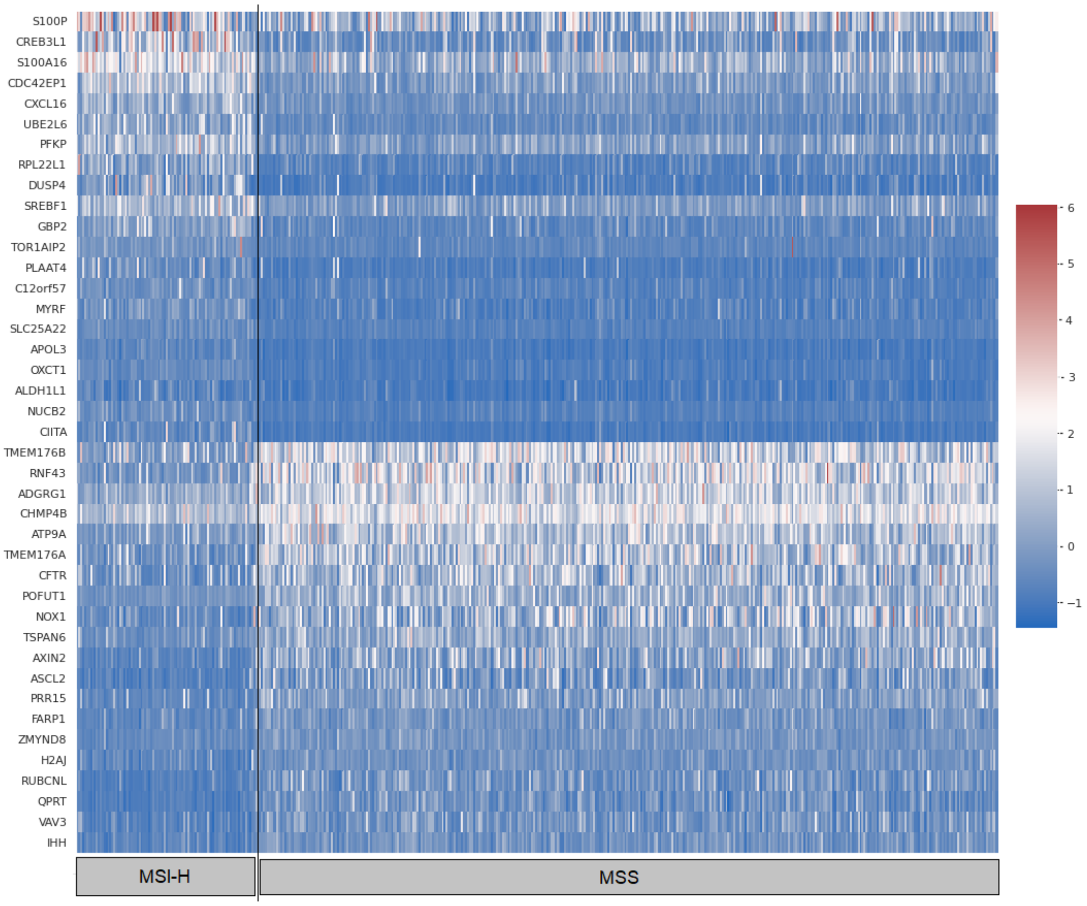

Figure 5.5: Heatmap of the most influential genes selected by DR + MRMR with the 2 class COAD dataset.

5.2.4 Dimensionality Reduction and ANOVA

The heatmap of the most relevant features with the combined method of DR and ANOVA, using the STAD dataset, is shown in figure 5.6. Genes *AGR2*, *SREBF1* and *S100P* have more expression in MSI-H samples, while genes *RPS6*, *ATP5F1B*, *BF3E* and *ATP5F1A* are more expressed in MSS samples.

AGR2 and *RPS6* values highlight themselves when compared to all the other genes.

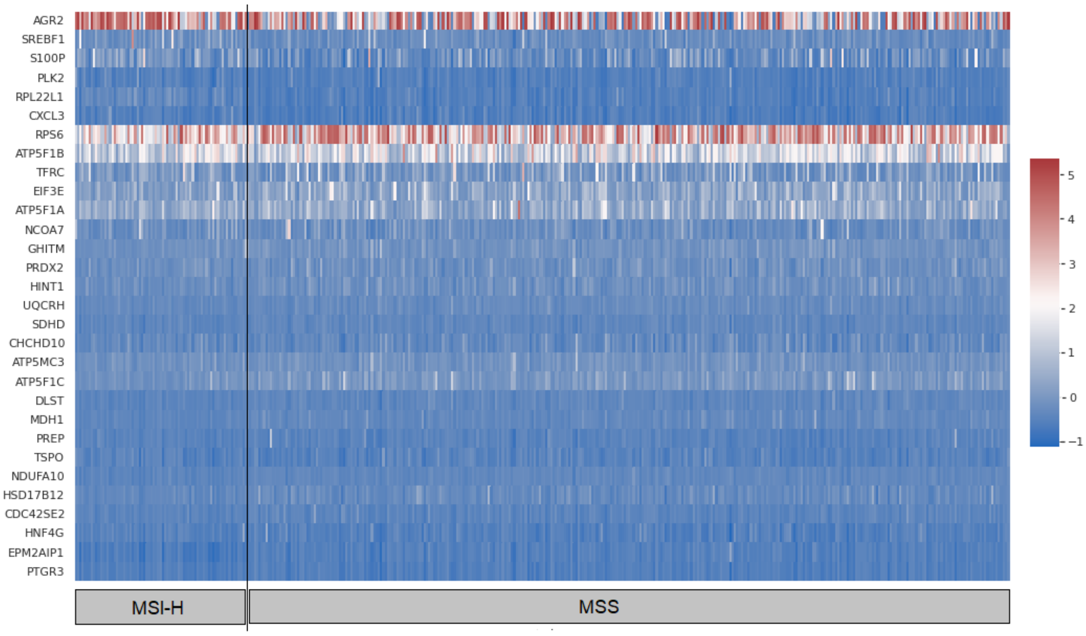

Figure 5.6: Heatmap of the most influential genes selected by DR + ANOVA in the 2 class STAD dataset.

This method's runtime was 16 minutes with COAD and UCEC, and 17 minutes with STAD.

5.2.5 15 Feature Set with T-Test

Figure 5.7 exhibits the heatmap of the genes in the 15-feature set with the multiclass UCEC dataset. It is clear that gene *RPL22L1* is more expressed in the MSI-H samples, while genes *DDX27*, *EPM2AIP1*, *MLH1*, *NOL4L* and *RTF2* are more expressed in the MSI-L and MSS samples.

Figure 5.8 presents the same gene expression levels, with the same dataset merging MSI-L and MSS samples, allowing the confirmation of the proximity between MSI-L and MSS samples. From this figure, it is possible to extract the same information as in the previous one.

5.3 Discussion

In section 5.1, an analysis of the results of different metrics is performed in the experiments made to predict MSI using RNA-seq data. That allowed us to make the following conclusions:

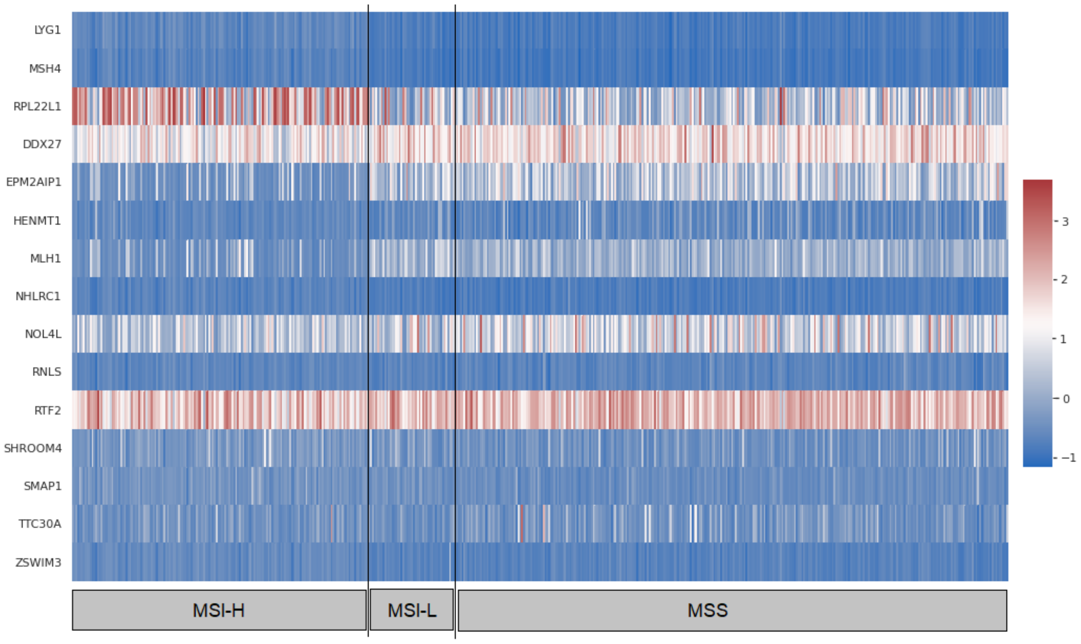

Figure 5.7: Heatmap of the 15 genes selected by the 15-feature set with the 3 class UCEC dataset.

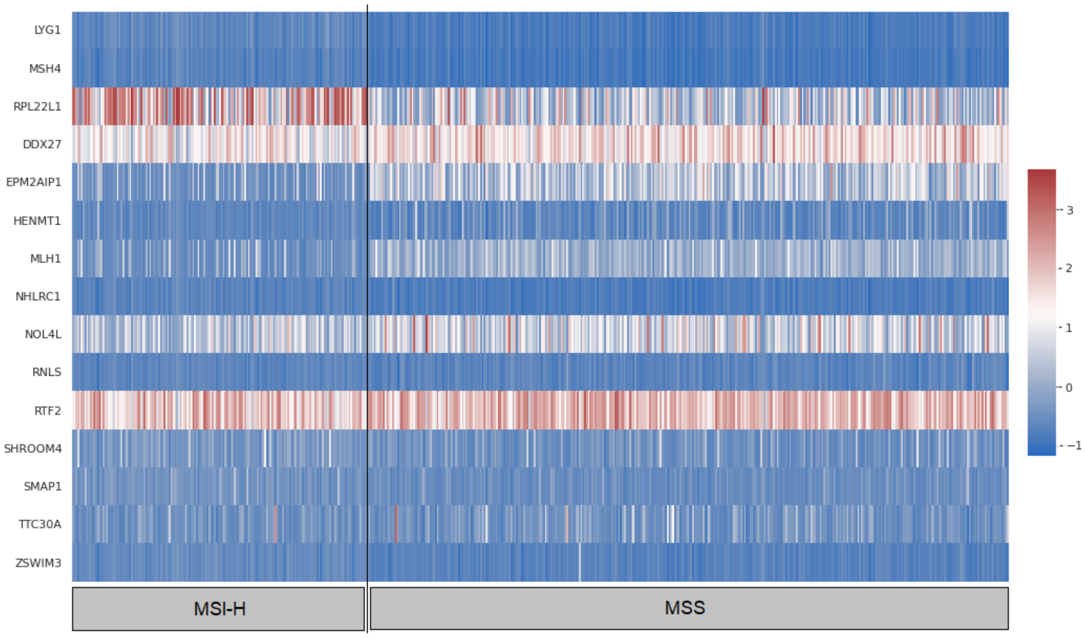

Figure 5.8: Heatmap of the 15 genes selected by the 15-feature set with the 2 class UCEC dataset.

- **The binary classification approach that merged MSI-L with MSS samples performed the best results**. With better AUC and accuracy scores than the other approaches, the merge of MSI-L and MSS samples confirmed the proximity of these classes, which was already referred to in some of the articles in the literature. With AUC and accuracy values almost always above 80%, this approach outperformed the multiclass and the binary classification merging MSI-L with MSI-H approaches.

- **MRMR was the feature selection with better overall performance between all approaches.** Analysing the AUC and accuracy scores from the MRMR and the DR + MRMR approaches, it is possible to conclude that this method performed better than ANOVA, with or without DR and the 15-feature set.

- **COAD dataset performed better than the other datasets when MSI-L was merged with MSS.** The results of the experiments show that the ones with the COAD dataset obtained better results than those with STAD or UCEC datasets in the same conditions as the experiments from the literature. With accuracy in a range of 87-92% and an AUC in a range from 90 to 98%, it outperformed the accuracy, ranging from 86 to 89%, and the AUC, ranging from 87 to 95%, of STAD. The same happened with UCEC, with accuracy ranging from 82 to 88% and an AUC ranging from 86 to 94%.

- **UCEC performed better in the multiclass and binary classification that merged MSI-L with MSI-H approaches.** While in the same conditions of the literature, COAD prevailed, on the others it was UCEC that stood, but its values were below the ones from the approach with better results. That can be explained by the conditions of the dataset, with less than 10% of the samples labelled as MSI-L, the intermediate class.

- **With 98.44% of AUC and 92.67% of accuracy, the binary classification MLP model that merges MSI-L with MSS, using the COAD dataset, was the best experiment.**

- The obtained results agree with the ones from [19]. The values in this study for COAD, STAD and UCEC with the KNN model in the binary approach that merged MSI-L with MSS are slightly lower than the ones from the previous study. The minor difference in the results can be explained by the data preparation and the tuning of the KNN model. Despite this, the results can validate the importance and values of Li's study.

- **The Dimensionality Reduction technique proved to be beneficial in analysing gene behaviour.** The feature analysis heatmaps with the DR technique comprehend genes with much more variability and expression levels between groups of samples. That shows this technique can be advantageous to analyse with more detail which genes are more influenced by MSI.

- **Genes *RPL22L1*, *AGR2*, *CREB3L1*, *DUSP4*, *CDC42EP1*, *UBE2L6*, *CXCL16* and *S100A16* are more expressed in MSI patients.** In, at least, two heatmaps, these genes were more expressed in MSI-H samples than MSI-L or MSS samples. According to NIH[1], most of these genes are related to the endoplasmatic reticulum, ribosome and cell regulation functions.

- **Genes *MLH1*, *EPM2AIP1*, *RNF43*, *PRR15*, *TMEM176A*, *CFTR*, *AXIN2* and *ADGRG1* are less expressed in MSI patients.** In, at least, two heatmaps, these genes were less expressed in MSI-H samples than MSI-L or MSS samples. An extensive set of genes was observed in this condition, most of them once. The first gene of this set is related to DNA MMR functions, as shown in figure 2.3. This result confirms the loss of MMR activity in patients with MSI. Other genes of this group are also associated with defective MMR and colon, stomach and endometrial cancers by NIH.

5.4 Summary

117 different experiments were done to predict MSI using RNAseq data. With a binary classification MLP, using the COAD dataset merging MSI-L with MSS, the best results were obtained with an AUC of 98.44% and an accuracy of 92.67%. In the feature selection methods analysis, different genes revealed different expression levels in the three classes of samples, with the DR technique showing the capacity to find those genes.

[1] https://www.ncbi.nlm.nih.gov/, last accessed 10/09/2022.

Chapter 6

Conclusions

Since cancer is a high mortality disease causing millions of deaths yearly, it is essential to understand how to prevent the disease instead of detecting to treating it. The analysis of Microsatellite Instability can be helpful in this goal, allowing early detection of cancer and the capacity to apply immunotherapy as an effective way of treatment, causing less pain and effects to patients.

Different studies are being realised to understand how MS can be a predictor of several cancers. Most of them rely on ML models but not DL models. The feature selection analysis becomes fundamental to understanding which genes should be considered to predict the MS condition. These significant genetic variations can become relevant biomarkers to help the development of a program capable of being used in the medical industry in MSI prediction.

This work was essential to compare the effectiveness of different ML and DL models with different feature selection methods, allowing researchers to follow distinct paths in the future confidently. The MLP models reached the best results, confirming that DL can produce better results than other ML models. With the COAD dataset, in the binary classification approach that merged MSI-L with MSS, MLP reached 98.44% of AUC and 92.67% of accuracy with the DR + MRMR feature selection method in the best experiment of this work. The analysed studies do not mention the use of DL models to predict MSI, and so that is a gap this study can help to start close. However, this is just a small step in comprehending how DL can better predict MSI.

On the other hand, the analysis of different approaches relatively to MS classification data also allows us to confirm the proximity between MSI-L and MSS samples since merging these groups produced better results than when MSI-L was merged with MSI-H, or the three classes were analysed separately. With the merge of MSI-L with MSI-H, the best result came with MLP, UCEC dataset and MRMR feature selection method, reaching 88.50% AUC and 82.92% accuracy. In the experiments with the three classes, the best result came with UCEC, MLP and MRMR, extracting features with MRMR, reaching an AUC of 85.28% and accuracy of 58.48%. With STAD and MRMR, that model reaches a lower AUC of 81.02% but a higher accuracy of 63.16%.

The feature selection highlighted some genes already referenced in the literature, such as 'ENSG00000076242.13', with great importance in the DNA MMR process. This also confirms the background concepts relative to Microsatellite Instability and Mismatch Repair. However, this field needs more research in the future to allow the validation of distinct genes as MSI biomarkers, which can potentiate the creation of validated software to allow clinicians to predict MSI quickly in laboratories and the medical environment.

The results of this study are promising and allow the definition of distinct paths of research in the near future, with some of them being listed below:

- **Exploration of other DL Models.** Since the DL model analysed performed better than the other ML models studied, one of the paths the following studies can follow is to use other DL models to predict MSI and compare their results with the ones obtained from the MLP.
- **Generation of a feature set based on this and previous studies results.** Since most of the studies use distinct feature selection methods and generate different sets of selected features, even if some of them are similar, it will be essential to understand the relationship between most of these genes to create a genetic group that allows MSI detection in cancer patients.
- **Merge of datasets from distinct cancers.** Most of the studies use datasets from different cancer types separately since each has its characteristics. Join some of them to understand if the cancer type influences the MSI prediction is a path should be followed shortly since it can bring important discoveries about the topic.
- **Combine RNAseq with other types of data.** RNAseq is proving itself a great predictor of MSI singly. Combining this type of data with other types will indeed allow the detection of relevant biological patterns that can improve the prediction of MSI in cancer patients even more.

References

[1] C. Richard Boland and Ajay Goel. Microsatellite instability in colorectal cancer. *Gastroenterology*, 138:2073, 2010.

[2] Silvia Cascianelli, Ivan Molineris, Claudio Isella, Marco Masseroli, and Enzo Medico. Machine learning for rna sequencing-based intrinsic subtyping of breast cancer. *Scientific Reports*, 10:14071, 12 2020.

[3] Patrick Danaher, Sarah Warren, SuFey Ong, Nathan Elliott, Alessandra Cesano, and Sean Ferree. A gene expression assay for simultaneous measurement of microsatellite instability and anti-tumor immune activity. *Journal for ImmunoTherapy of Cancer*, 7:15, 12 2019.

[4] Elise J. Devlin, Linley A. Denson, and Hayley S. Whitford. Cancer treatment side effects: A meta-analysis of the relationship between response expectancies and experience. *Journal of Pain and Symptom Management*, 54:245–258.e2, 8 2017.

[5] Madeline Drexler. The cancer miracle isn't a cure. it's prevention. | harvard public health magazine | harvard t.h. chan school of public health. Available at: https://www.hsph.harvard.edu/magazine/magazine_article/the-cancer-miracle-isnt-a-cure-its-prevention/. Last accessed: 2022/02/01.

[6] Filippo Pietrantonio Elisabetta Puliga, Simona Corso and Silvia Giordano. Microsatellite instability in gastric cancer: Between lights and shadows. *Cancer Treatment Reviews*, 95:102175, 4 2021.

[7] K. Esfahani, L. Roudaia, N. Buhlaiga, S.V. Del Rincon, N. Papneja, and W.H. Miller. A review of cancer immunotherapy: From the past, to the present, to the future. *Current Oncology*, 27:87–97, 4 2020.

[8] T. S. Gerashchenko, E. V. Denisov, N. V. Litviakov, M. V. Zavyalova, S. V. Vtorushin, M. M. Tsyganov, V. M. Perelmuter, and N. V. Cherdyntseva. Intratumor heterogeneity: Nature and biological significance. *Biochemistry (Moscow)*, 78:1201–1215, 11 2013.

[9] Timothy P Hanna, Will D King, Stephane Thibodeau, Matthew Jalink, Gregory A Paulin, Elizabeth Harvey-Jones, Dylan E O'Sullivan, Christopher M Booth, Richard Sullivan, and Ajay Aggarwal. Mortality due to cancer treatment delay: systematic review and meta-analysis. *BMJ*, page m4087, 11 2020.

[10] Nigel Hawkes. Cancer survival data emphasise importance of early diagnosis. *BMJ*, 364:l408, 1 2019.

[11] Lindsey A. Hildebrand, Colin J. Pierce, Michael Dennis, Munizay Paracha, and Asaf Maoz. Artificial intelligence for histology-based detection of microsatellite instability and prediction of response to immunotherapy in colorectal cancer. *Cancers 2021, Vol. 13, Page 391*, 13:391, 1 2021.

[12] Willy Hugo, Jesse M. Zaretsky, Lu Sun, Chunying Song, Blanca Homet Moreno, Siwen Hu-Lieskovan, Beata Berent-Maoz, Jia Pang, Bartosz Chmielowski, Grace Cherry, Elizabeth Seja, Shirley Lomeli, Xiangju Kong, Mark C. Kelley, Jeffrey A. Sosman, Douglas B. Johnson, Antoni Ribas, and Roger S. Lo. Genomic and transcriptomic features of response to anti-pd-1 therapy in metastatic melanoma. *Cell*, 165:35–44, 3 2016.

[13] National Cancer Institute. Types of cancer treatment. Available at: `https://www.cancer.gov/about-cancer/treatment/types`. Last accessed: 2022/02/25.

[14] National Cancer Institute. Chemotherapy to treat cancer. Available at: `https://www.cancer.gov/about-cancer/treatment/types/chemotherapy`, 2015. Last accessed: 2022/02/25.

[15] National Cancer Institute. Radiation therapy for cancer. Available at: `https://www.cancer.gov/about-cancer/treatment/types/radiation-therapy`, 2019. Last accessed: 2022/02/25.

[16] National Cancer Institute. What is cancer? `https://www.cancer.gov/about-cancer/understanding/what-is-cancer`, 2021. Last accessed: 2022/02/04.

[17] Will et. al. Koehrsen. feature-selector. `https://github.com/WillKoehrsen/feature-selector`, 2022.

[18] Kai Li, Haiqing Luo, Lianfang Huang, Hui Luo, and Xiao Zhu. Microsatellite instability: A review of what the oncologist should know. *Cancer Cell International*, 20:1–13, 1 2020.

[19] Qiushi Feng Lin Li and Xiaosheng Wang. PreMSIm: An R package for predicting microsatellite instability from the expression profiling of a gene panel in cancer. *Computational and Structural Biotechnology Journal*, 18:668–675, 1 2020.

[20] Laurence Loewe. Genetic mutation. Available at: `https://www.nature.com/scitable/topicpage/genetic-mutation-1127/`, 2008. Last accessed: 2022/02/18.

[21] Ruairi J. Mackenzie. Dna vs. rna – 5 key differences and comparison. Available at: `https://www.technologynetworks.com/genomics/lists/what-are-the-key-differences-between-dna-and-rna-296719`, 2021. Last accessed: 2022/02/21.

[22] Andrija Matak, Pooja Lahiri, Ethan Ford, Daniela Pabst, Karl Kashofer, Dimitris Stellas, Dimitris Thanos, and Kurt Zatloukal. Stochastic phenotype switching leads to intratumor heterogeneity in human liver cancer. *Hepatology*, 68:933–948, 9 2018.

[23] Usha Moorthy and Usha Devi Gandhi. A novel optimal feature selection technique for medical data classification using anova based whale optimization. *Journal of Ambient Intelligence and Humanized Computing*, 12:3527–3538, 3 2021.

[24] The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487:330–337, 7 2012.

[25] The Cancer Genome Atlas Network. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, 513:202–209, 9 2014.

[26] Eindhoven University of Technology. Machine learning helps in predicting when immunotherapy will be effective. *ScienceDaily*, 6 2021.

[27] Anna Pačínková and Vlad Popovici. Cross-platform data analysis reveals a generic gene expression signature for microsatellite instability in colorectal cancer. *BioMed Research International*, 2019:1–9, 3 2019.

[28] Hanchuan Peng, Fuhui Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

[29] Mi-Kyoung Seo, Hyundeok Kang, and Sangwoo Kim. Tumor microenvironment-aware, single-transcriptome prediction of microsatellite instability in colorectal cancer using meta-analysis. *Scientific Reports*, 12:6283, 12 2022.

[30] American Cancer Society. Goals of chemotherapy | how is chemotherapy given? Available at: https://www.cancer.org/treatment/treatments-and-side-effects/treatment-types/chemotherapy/how-is-chemotherapy-used-to-treat-cancer.html, 2019. Last accessed: 2022/02/25.

[31] American Cancer Society. How does chemo work? | Types of Chemotherapy. Available at: https://www.cancer.org/treatment/treatments-and-side-effects/treatment-types/chemotherapy/how-chemotherapy-drugs-work.html, 2019. Last accessed: 2022/02/25.

[32] American Cancer Society. How radiation therapy is used to treat cancer. Available at: https://www.cancer.org/treatment/treatments-and-side-effects/treatment-types/radiation/basics.html, 2019. Last accessed: 2022/02/25.

[33] American Cancer Society. What is cancer? https://www.cancer.org/treatment/understanding-your-diagnosis/what-is-cancer.html, 2020. Last accessed: 2022/02/04.

[34] Maksim Sorokin, Elizaveta Rabushko, Victor Efimov, Elena Poddubskaya, Marina Sekacheva, Alexander Simonov, Daniil Nikitin, Aleksey Drobyshev, Maria Suntsova, and Anton Buzdin. Experimental and meta-analytic validation of rna sequencing signatures for predicting status of microsatellite instability. *Frontiers in Molecular Biosciences*, 8, 11 2021.

[35] Michael R. Stratton, Peter J. Campbell, and P. Andrew Futreal. The cancer genome. *Nature*, 458:719–724, 4 2009.

[36] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71:209–249, 5 2021.

[37] C Lee Ventola. Cancer immunotherapy, part 1: Current strategies and agents. *P & T : a peer-reviewed journal for formulary management*, 42:375–383, 6 2017.

[38] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10:57–63, 1 2009.

[39] H. Richard Winn. *Genetic Origins of Brain Tumors*. Elsevier Health Sciences, 7th edition, 2017.

[40] Santiago Ramón y Cajal, Marta Sesé, Claudia Capdevila, Trond Aasen, Leticia De Mattos-Arruda, Salvador J. Diaz-Cano, Javier Hernández-Losa, and Josep Castellví. Clinical implications of intratumor heterogeneity: challenges and opportunities. *Journal of Molecular Medicine*, 98:161–177, 2 2020.

[41] Yuanyuan Zhang and Zemin Zhang. The history and advances in cancer immunotherapy: understanding the characteristics of tumor-infiltrating immune cells and their therapeutic implications. *Cellular & Molecular Immunology*, 17:807–821, 8 2020.

[42] Shuangshuang Zeng Xinxin Ren Yuanliang Yan Zhijie Xu, Xiang Wang and Zhicheng Gong. Applying artificial intelligence for cancer immunotherapy. *Acta Pharmaceutica Sinica B*, 11:3393–3405, 11 2021.

[43] Óscar Lapuente-Santana, Maisa van Genderen, Peter A.J. Hilbers, Francesca Finotello, and Federica Eduati. Interpretable systems biomarkers predict response to immune-checkpoint inhibitors. *Patterns*, 2:100293, 8 2021.

Appendix A

Results Table

Table A.1: List of the all the results for the experiments with COAD dataset.

Approach	Algorithm	Feature Selection Method	AUC (%)	Balanced Accuracy (%)	Specificity (%)	Precision (%)	Sensitivity (%)	F1 Score (%)
Multiclass	**Multi-Layer Perceptron**	**MRMR**	**78.38 +/- 3.59**	62.32 +/- 5.30	64.61 +/- 8.30	**69.75 +/- 4.01**	58.65 +/- 4.87	61.49 +/- 4.33
		ANOVA	77.86 +/- 2.71	**62.34 +/- 4.84**	63.91 +/- 6.51	69.24 +/- 3.15	**61.26 +/- 5.16**	**63.15 +/- 4.53**
		15-Feature Set	76.88 +/- 2.85	61.54 +/- 4.67	**68.62 +/- 7.38**	68.12 +/- 4.78	54.17 +/- 5.67	56.11 +/- 6.64
	Random Forest	**MRMR**	**71.33 +/- 3.76**	**55.58 +/- 5.37**	64.76 +/- 12.44	61.66 +/- 10.30	**46.13 +/- 10.66**	**46.09 +/- 12.30**
		ANOVA	70.95 +/- 3.38	55.21 +/- 4.44	**70.34 +/- 10.71**	61.14 +/- 19.79	40.08 +/- 15.43	32.60 +/- 16.84
		15-Feature Set	68.59 +/- 3.63	54.23 +/- 5.28	66.05 +/- 11.60	**63.37 +/- 10.64**	42.59 +/- 10.51	42.32 +/- 11.98
	K-Nearest Neighbours	**MRMR**	**73.28 +/- 3.46**	**59.30 +/- 5.72**	**68.82 +/- 7.09**	**67.06 +/- 5.47**	49.77 +/- 5.63	51.80 +/- 6.06
		ANOVA	71.17 +/- 4.13	58.54 +/- 5.44	65.10 +/- 7.53	66.37 +/- 4.57	**51.98 +/- 4.53**	**54.52 +/- 4.49**
		15-Feature Set	72.07 +/- 3.52	57.26 +/- 5.02	64.85 +/- 6.84	63.60 +/- 4.22	49.66 +/- 4.32	51.59 +/- 4.18
Binary Classification **(MSI-L merged with MSI-H)**	**Multi-Layer Perceptron**	**DR + MRMR**	78.50 +/- 3.85	**72.30 +/- 4.00**	68.79 +/- 4.50	74.05 +/- 3.54	74.00 +/- 3.63	73.71 +/- 3.57
		DR + ANOVA	78.16 +/- 4.52	71.65 +/- 3.88	**69.76 +/- 4.53**	**74.66 +/- 3.59**	74.57 +/- 3.53	**74.34 +/- 3.49**
		MRMR	**79.24 +/- 4.30**	71.79 +/- 3.99	69.19 +/- 4.99	73.07 +/- 3.92	72.69 +/- 3.84	72.70 +/- 3.82
		ANOVA	76.69 +/- 4.66	71.45 +/- 4.24	68.13 +/- 4.85	74.54 +/- 4.11	**74.76 +/- 3.98**	74.24 +/- 4.01
		15-Feature Set	73.79 +/- 4.82	69.08 +/- 4.65	66.39 +/- 5.35	71.86 +/- 4.42	71.77 +/- 4.56	71.46 +/- 4.46
	Random Forest	**DR + MRMR**	70.14 +/- 5.17	67.45 +/- 4.06	63.43 +/- 5.30	69.53 +/- 5.52	68.78 +/- 6.40	68.36 +/- 5.95
		DR + ANOVA	**72.52 +/- 4.69**	**68.60 +/- 4.43**	64.48 +/- 4.00	71.71 +/- 5.30	70.46 +/- 5.93	69.81 +/- 5.22
		MRMR	72.41 +/- 4.85	67.85 +/- 4.28	**65.41 +/- 5.16**	71.47 +/- 4.13	70.29 +/- 5.50	69.75 +/- 5.16
		ANOVA	69.57 +/- 5.17	66.53 +/- 4.68	63.41 +/- 4.53	69.96 +/- 5.06	69.64 +/- 5.71	69.15 +/- 5.25
		15-Feature Set	71.41 +/- 6.05	67.93 +/- 5.19	63.39 +/- 5.49	**73.04 +/- 6.01**	**72.46 +/- 5.88**	**71.27 +/- 5.50**
	K-Nearest Neighbours	**DR + MRMR**	75.09 +/- 4.79	67.50 +/- 4.29	68.72 +/- 4.83	**70.85 +/- 4.17**	**68.95 +/- 4.63**	**69.34 +/- 4.52**
		DR + ANOVA	**76.21 +/- 4.21**	**70.24 +/- 4.01**	**69.27 +/- 4.31**	70.73 +/- 3.63	67.89 +/- 3.80	68.38 +/- 3.71
		MRMR	72.11 +/- 4.76	65.61 +/- 4.37	65.88 +/- 4.83	68.02 +/- 4.08	65.35 +/- 4.45	65.81 +/- 4.33
		ANOVA	75.47 +/- 4.43	68.68 +/- 4.01	69.20 +/- 4.25	70.75 +/- 3.68	68.15 +/- 4.07	68.63 +/- 3.99
		15-Feature Set	70.47 +/- 4.72	64.93 +/- 4.94	63.42 +/- 5.34	67.41 +/- 4.63	66.44 +/- 5.06	66.60 +/- 4.89
Binary Classification **(MSI-L merged with MSS)**	**Multi-Layer Perceptron**	**DR + MRMR**	**98.44 +/- 1.20**	**92.67 +/- 3.07**	**93.91 +/- 4.27**	**94.40 +/- 1.82**	93.24 +/- 2.42	**93.52 +/- 2.25**
		DR + ANOVA	97.53 +/- 1.61	92.49 +/- 3.56	92.52 +/- 5.19	93.73 +/- 2.12	92.44 +/- 2.77	92.76 +/- 2.57
		MRMR	97.62 +/- 1.56	92.28 +/- 3.92	89.33 +/- 6.01	93.85 +/- 2.49	**93.31 +/- 2.78**	93.45 +/- 2.67
		ANOVA	97.12 +/- 1.54	91.25 +/- 3.25	91.03 +/- 4.79	93.00 +/- 2.04	91.47 +/- 2.89	91.86 +/- 2.64
		15-Feature Set	96.19 +/- 2.00	89.44 +/- 3.70	89.03 +/- 5.96	91.74 +/- 1.97	89.85 +/- 2.56	90.34 +/- 2.33
	Random Forest	**DR + MRMR**	90.53 +/- 5.32	87.78 +/- 5.02	83.92 +/- 7.58	91.22 +/- 2.60	90.21 +/- 3.15	90.47 +/- 2.92
		DR + ANOVA	89.78 +/- 5.33	87.13 +/- 3.89	83.07 +/- 7.13	90.58 +/- 2.95	89.49 +/- 3.42	89.82 +/- 3.20
		MRMR	**90.97 +/- 4.60**	86.86 +/- 4.67	84.78 +/- 6.15	90.88 +/- 2.71	88.81 +/- 4.12	89.62 +/- 3.25
		ANOVA	90.83 +/- 4.52	**87.86 +/- 4.71**	**85.25 +/- 7.36**	**91.53 +/- 2.66**	**90.46 +/- 3.15**	**90.75 +/- 2.94**
		15-Feature Set	**90.97 +/- 4.60**	85.17 +/- 3.45	84.55 +/- 6.84	90.52 +/- 2.74	88.55 +/- 4.23	89.07 +/- 3.76
	K-Nearest Neighbours	**DR + MRMR**	95.11 +/- 2.51	87.85 +/- 4.24	88.28 +/- 7.25	90.67 +/- 2.39	87.73 +/- 3.27	88.46 +/- 2.94
		DR + ANOVA	95.32 +/- 2.30	89.91 +/- 3.79	88.38 +/- 6.58	89.99 +/- 1.98	86.00 +/- 2.88	86.98 +/- 2.53
		MRMR	**95.77 +/- 2.02**	**91.63 +/- 2.89**	**93.03 +/- 4.01**	**92.73 +/- 1.88**	**90.23 +/- 3.01**	**90.82 +/- 2.70**
		ANOVA	93.48 +/- 3.20	88.70 +/- 4.08	87.92 +/- 6.13	91.32 +/- 2.31	89.47 +/- 2.84	89.98 +/- 2.62
		15-Feature Set	93.91 +/- 2.64	88.47 +/- 3.47	88.80 +/- 5.20	90.94 +/- 2.03	88.15 +/- 3.21	88.87 +/- 2.84

Table A.2: List of the all the results for the experiments with STAD dataset.

Approach	Algorithm	Feature Selection Method	AUC (%)	Balanced Accuracy (%)	Specificity (%)	Precision (%)	Sensitivity (%)	F1 Score (%)
Multiclass	**Multi-Layer Perceptron**	**MRMR**	**81.02 +/- 3.89**	**63.16 +/- 5.72**	**65.40 +/- 9.30**	**74.17 +/- 3.97**	**63.49 +/- 5.42**	**66.54 +/- 4.84**
		ANOVA	72.07 +/- 5.49	55.99 +/- 6.95	53.09 +/- 9.91	66.17 +/- 4.96	55.98 +/- 6.76	58.96 +/- 6.29
		15-Feature Set	75.81 +/- 4.56	58.01 +/- 6.32	63.29 +/- 10.25	68.94 +/- 5.26	53.78 +/- 8.63	59.92 +/- 8.08
	Random Forest	**MRMR**	**72.35 +/- 4.79**	**56.68 +/- 5.25**	58.87 +/- 15.33	**71.77 +/- 6.01**	**52.49 +/- 10.82**	**55.75 +/- 10.41**
		ANOVA	70.14 +/- 4.65	54.14 +/- 5.64	**59.67 +/- 11.58**	68.79 +/- 4.15	49.24 +/- 8.79	53.14 +/- 7.61
		15-Feature Set	71.50 +/- 4.58	54.59 +/- 5.97	57.72 +/- 11.15	69.10 +/- 5.27	51.85 +/- 9.19	55.64 +/- 8.45
	K-Nearest Neighbours	**MRMR**	**72.43 +/- 3.65**	56.68 +/- 4.85	59.11 +/- 6.46	**68.31 +/- 3.27**	**54.24 +/- 5.46**	**57.70 +/- 5.10**
		ANOVA	68.96 +/- 4.99	51.70 +/- 6.95	52.85 +/- 9.88	65.19 +/- 4.46	50.56 +/- 5.48	53.88 +/- 5.10
		15-Feature Set	71.81 +/- 4.91	**57.51 +/- 7.24**	**65.29 +/- 10.25**	66.82 +/- 5.86	49.73 +/- 5.88	53.06 +/- 5.76
Binary Classification (MSI-L merged with MSI-H)	**Multi-Layer Perceptron**	**DR + MRMR**	**81.29 +/- 4.76**	**74.09 +/- 4.60**	**72.53 +/- 5.80**	**77.25 +/- 4.26**	**76.15 +/- 4.72**	**76.35 +/- 4.52**
		DR + ANOVA	73.39 +/- 5.40	69.25 +/- 3.76	65.39 +/- 6.32	70.64 +/- 4.46	68.59 +/- 4.85	69.08 +/- 4.67
		MRMR	80.08 +/- 5.25	74.01 +/- 5.23	71.35 +/- 6.14	76.48 +/- 4.49	75.51 +/- 4.85	75.69 +/- 4.68
		ANOVA	74.14 +/- 5.44	68.79 +/- 4.49	64.85 +/- 5.55	70.87 +/- 4.01	69.49 +/- 4.39	69.82 +/- 4.19
		15-Feature Set	74.88 +/- 5.62	69.75 +/- 4.73	65.87 +/- 6.01	72.71 +/- 4.23	72.02 +/- 4.37	72.07 +/- 4.20
	Random Forest	**DR + MRMR**	74.97 +/- 5.48	**70.33 +/- 5.18**	67.77 +/- 6.05	73.41 +/- 5.18	70.90 +/- 6.81	71.15 +/- 6.27
		DR + ANOVA	**75.31 +/- 5.72**	70.02 +/- 5.48	**67.82 +/- 6.31**	**74.99 +/- 5.01**	**73.85 +/- 6.13**	**73.76 +/- 5.68**
		MRMR	74.09 +/- 5.52	68.99 +/- 5.69	67.56 +/- 6.38	73.08 +/- 5.03	70.41 +/- 7.10	70.65 +/- 6.62
		ANOVA	72.08 +/- 5.50	67.10 +/- 4.49	61.22 +/- 6.61	72.58 +/- 5.16	70.78 +/- 6.92	69.81 +/- 6.03
		15-Feature Set	72.17 +/- 4.59	68.19 +/- 4.61	65.26 +/- 5.57	73.17 +/- 5.04	71.41 +/- 6.44	71.27 +/- 5.83
	K-Nearest Neighbours	**DR + MRMR**	**76.87 +/- 5.57**	**70.26 +/- 5.18**	70.06 +/- 5.55	**72.62 +/- 3.96**	**68.12 +/- 4.36**	**68.92 +/- 4.19**
		DR + ANOVA	75.20 +/- 5.33	69.03 +/- 5.88	**70.40 +/- 5.51**	72.43 +/- 4.22	66.83 +/- 4.97	67.69 +/- 4.82
		MRMR	73.43 +/- 4.66	66.56 +/- 4.62	65.32 +/- 5.27	70.24 +/- 3.88	67.76 +/- 4.46	68.36 +/- 4.20
		ANOVA	71.02 +/- 4.85	65.32 +/- 5.00	63.88 +/- 6.04	69.09 +/- 4.33	66.76 +/- 4.50	67.40 +/- 4.35
		15-Feature Set	72.46 +/- 4.91	65.61 +/- 4.88	64.07 +/- 5.77	69.47 +/- 4.21	67.22 +/- 4.84	67.77 +/- 4.63
Binary Classification (MSI-L merged with MSS)	**Multi-Layer Perceptron**	**DR + MRMR**	91.80 +/- 5.63	87.00 +/- 5.01	81.17 +/- 9.52	91.21 +/- 3.13	90.76 +/- 3.19	90.85 +/- 3.15
		DR + ANOVA	89.82 +/- 5.05	85.65 +/- 5.51	78.28 +/- 9.34	90.06 +/- 3.35	89.44 +/- 3.81	89.58 +/- 3.61
		MRMR	93.06 +/- 4.41	87.88 +/- 4.70	83.61 +/- 7.45	**92.37 +/- 2.53**	**91.98 +/- 2.64**	**92.06 +/- 2.56**
		ANOVA	88.77 +/- 5.93	82.00 +/- 5.48	76.94 +/- 9.47	88.34 +/- 2.92	87.12 +/- 3.00	87.49 +/- 2.87
		15-Feature Set	**95.78 +/- 2.07**	**89.17 +/- 3.93**	**86.81 +/- 6.20**	92.10 +/- 2.22	91.00 +/- 2.67	91.32 +/- 2.49
	Random Forest	**DR + MRMR**	87.08 +/- 7.09	85.66 +/- 5.41	78.80 +/- 1.24	90.25 +/- 2.71	88.98 +/- 2.91	89.12 +/- 2.71
		DR + ANOVA	**87.98 +/- 7.87**	85.42 +/- 6.18	81.30 +/- 10.30	91.17 +/- 3.03	90.44 +/- 3.15	90.57 +/- 3.08
		MRMR	86.99 +/- 6.99	85.41 +/- 5.83	79.90 +/- 1.08	91.16 +/- 2.82	80.63 +/- 2.74	80.65 +/- 2.78
		ANOVA	87.78 +/- 7.23	**86.97 +/- 5.81**	**82.42 +/- 10.28**	**91.88 +/- 2.53**	**91.27 +/- 2.52**	**91.34 +/- 2.52**
		15-Feature Set	86.70 +/- 4.89	82.25 +/- 4.88	77.44 +/- 8.81	88.89 +/- 2.65	87.54 +/- 3.25	87.89 +/- 2.93
	K-Nearest Neighbours	**DR + MRMR**	86.59 +/- 4.98	83.15 +/- 5.10	79.91 +/- 8.42	85.62 +/- 2.67	77.76 +/- 4.11	79.81 +/- 3.49
		DR + ANOVA	79.11 +/- 5.78	79.27 +/- 5.83	71.02 +/- 10.29	82.33 +/- 3.31	74.54 +/- 3.89	76.83 +/- 3.36
		MRMR	92.56 +/- 4.60	87.33 +/- 4.32	88.35 +/- 7.30	90.24 +/- 2.23	86.32 +/- 2.99	87.30 +/- 2.66
		ANOVA	85.44 +/- 6.36	78.30 +/- 5.75	78.53 +/- 9.00	85.34 +/- 3.10	78.07 +/- 4.87	80.04 +/- 4.18
		15-Feature Set	**94.14 +/- 2.79**	**89.75 +/- 3.38**	**89.23 +/- 5.37**	**92.16 +/- 1.97**	**90.27 +/- 2.93**	**90.77 +/- 2.62**

Table A.3: List of the all the results for the experiments with UCEC dataset.

Approach	Algorithm	Feature Selection Method	AUC (%)	Balanced Accuracy (%)	Specificity (%)	Precision (%)	Sensitivity (%)	F1 Score (%)
Multiclass	**Multi-Layer Perceptron**	**MRMR**	**85.28 +/- 2.63**	**58.48 +/- 5.34**	50.55 +/- 8.33	**77.31 +/- 3.45**	**63.64 +/- 6.06**	**68.31 +/- 5.40**
		ANOVA	82.71 +/- 2.80	56.79 +/- 5.55	50.00 +/- 9.61	74.81 +/- 3.21	63.58 +/- 3.81	67.62 +/- 3.20
		15-Feature Set	80.24 +/- 3.32	54.20 +/- 5.03	**53.92 +/- 8.03**	72.28 +/- 4.08	54.48 +/- 6.51	58.08 +/- 7.26
	Random Forest	**MRMR**	**79.48 +/- 4.00**	**57.74 +/- 6.39**	**53.66 +/- 11.30**	**77.12 +/- 4.60**	**61.81 +/- 7.57**	**66.68 +/- 6.09**
		ANOVA	79.44 +/- 3.95	55.98 +/- 4.90	50.97 +/- 11.43	75.52 +/- 4.52	60.99 +/- 8.45	65.18 +/- 7.12
		15-Feature Set	76.68 +/- 3.70	56.89 +/- 5.42	59.43 +/- 11.69	74.49 +/- 4.70	54.34 +/- 8.77	59.16 +/- 8.71
	K-Nearest Neighbours	**MRMR**	79.54 +/- 3.41	54.26 +/- 6.12	50.21 +/- 9.66	72.30 +/- 3.69	58.32 +/- 4.90	62.69 +/- 4.44
		ANOVA	77.21 +/- 2.92	49.59 +/- 7.16	49.13 +/- 11.09	71.16 +/- 4.24	50.06 +/- 5.15	55.42 +/- 4.96
		15-Feature Set	**81.13 +/- 3.59**	**56.38 +/- 6.01**	**53.48 +/- 10.04**	**75.24 +/- 3.90**	**59.28 +/- 4.54**	**64.25 +/- 4.18**
Binary Classification (MSI-L merged with MSI-H)	**Multi-Layer Perceptron**	**DR + MRMR**	**88.57 +/- 3.55**	82.79 +/- 3.90	**81.36 +/- 4.34**	82.76 +/- 3.99	82.55 +/- 3.98	82.52 +/- 3.97
		DR + ANOVA	87.69 +/- 3.29	82.21 +/- 3.47	80.04 +/- 3.89	81.82 +/- 3.65	81.66 +/- 3.67	81.59 +/- 3.65
		MRMR	88.50 +/- 3.50	**82.92 +/- 3.88**	80.95 +/- 3.67	**83.36 +/- 2.92**	**83.14 +/- 2.92**	**82.99 +/- 2.98**
		ANOVA	87.61 +/- 3.50	81.55 +/- 3.40	80.71 +/- 3.68	82.06 +/- 3.23	81.81 +/- 3.33	81.79 +/- 3.34
		15-Feature Set	81.49 +/- 4.08	76.07 +/- 4.23	73.95 +/- 4.76	78.47 +/- 4.64	78.00 +/- 4.30	77.52 +/- 4.34
	Random Forest	**DR + MRMR**	**83.20 +/- 4.10**	76.25 +/- 4.79	76.18 +/- 4.38	78.61 +/- 4.13	77.73 +/- 4.71	77.55 +/- 4.72
		DR + ANOVA	82.64 +/- 4.17	77.00 +/- 3.79	76.01 +/- 4.73	78.49 +/- 4.10	77.89 +/- 3.99	77.71 +/- 3.99
		MRMR	82.76 +/- 4.33	76.86 +/- 4.28	75.68 +/- 4.47	78.86 +/- 4.43	77.96 +/- 4.54	77.70 +/- 4.44
		ANOVA	82.22 +/- 5.04	**77.27 +/- 4.45**	**76.25 +/- 4.92**	78.79 +/- 4.23	78.42 +/- 4.13	**78.24 +/- 4.17**
		15-Feature Set	80.25 +/- 4.51	76.44 +/- 4.19	74.23 +/- 4.53	**79.45 +/- 4.49**	**78.65 +/- 4.21**	78.07 +/- 4.21
	K-Nearest Neighbours	**DR + MRMR**	84.61 +/- 3.80	75.94 +/- 4.56	77.43 +/- 4.78	78.34 +/- 4.12	77.77 +/- 4.04	77.83 +/- 4.06
		DR + ANOVA	**85.28 +/- 4.37**	76.52 +/- 5.00	**78.03 +/- 4.97**	**79.61 +/- 4.30**	**79.28 +/- 4.24**	**79.23 +/- 4.25**
		MRMR	85.02 +/- 4.54	**77.79 +/- 4.37**	77.65 +/- 4.98	78.52 +/- 4.12	77.92 +/- 3.89	77.97 +/- 3.94
		ANOVA	83.81 +/- 4.48	77.19 +/- 4.54	76.39 +/- 5.00	78.11 +/- 4.25	78.00 +/- 4.17	77.94 +/- 4.24
		15-Feature Set	83.99 +/- 3.97	77.59 +/- 3.78	76.18 +/- 4.20	78.95 +/- 3.60	78.90 +/- 3.59	78.68 +/- 3.68
Binary Classification (MSI-L merged with MSS)	**Multi-Layer Perceptron**	**DR + MRMR**	94.41 +/- 2.38	87.50 +/- 3.44	86.17 +/- 4.44	89.09 +/- 2.95	88.72 +/- 3.18	88.77 +/- 3.12
		DR + ANOVA	92.86 +/- 2.25	85.54 +/- 3.40	83.02 +/- 4.14	86.23 +/- 2.82	85.52 +/- 3.18	85.65 +/- 3.06
		MRMR	**94.47 +/- 2.22**	**88.07 +/- 3.13**	**86.66 +/- 4.37**	**89.23 +/- 2.72**	**88.72 +/- 2.98**	**88.79 +/- 2.92**
		ANOVA	91.87 +/- 2.74	84.40 +/- 3.27	82.72 +/- 4.12	86.26 +/- 2.70	85.64 +/- 3.04	85.73 +/- 2.93
		15-Feature Set	87.86 +/- 4.20	82.13 +/- 3.55	80.63 +/- 5.30	83.97 +/- 3.44	82.95 +/- 3.67	83.15 +/- 3.56
	Random Forest	**DR + MRMR**	85.86 +/- 4.15	80.07 +/- 4.27	78.76 +/- 5.59	82.30 +/- 3.85	82.38 +/- 4.29	82.49 +/- 4.14
		DR + ANOVA	85.39 +/- 3.79	80.26 +/- 3.96	78.30 +/- 4.37	82.78 +/- 3.22	81.81 +/- 3.76	81.96 +/- 3.53
		MRMR	84.80 +/- 4.84	80.79 +/- 4.57	78.22 +/- 5.45	82.76 +/- 3.71	81.92 +/- 4.10	82.05 +/- 3.95
		ANOVA	86.10 +/- 4.44	80.84 +/- 4.62	78.47 +/- 5.48	84.10 +/- 4.38	83.31 +/- 4.68	83.32 +/- 4.45
		15-Feature Set	**86.53 +/- 3.99**	**82.33 +/- 3.62**	**79.66 +/- 5.00**	**85.30 +/- 2.93**	**82.49 +/- 4.14**	**84.91 +/- 3.04**
	K-Nearest Neighbours	**DR + MRMR**	87.82 +/- 3.47	80.18 +/- 4.12	80.01 +/- 4.70	80.84 +/- 3.28	77.01 +/- 3.78	77.72 +/- 3.61
		DR + ANOVA	87.25 +/- 3.38	78.04 +/- 3.49	78.27 +/- 5.20	79.95 +/- 3.32	76.59 +/- 3.46	77.25 +/- 3.33
		MRMR	88.49 +/- 3.19	79.53 +/- 4.45	79.76 +/- 5.42	81.10 +/- 3.75	78.11 +/- 4.09	78.73 +/- 3.95
		ANOVA	84.91 +/- 3.55	76.93 +/- 3.93	78.48 +/- 4.57	79.55 +/- 3.26	75.39 +/- 3.82	76.17 +/- 3.66
		15-Feature Set	**89.57 +/- 3.63**	**82.73 +/- 3.94**	**81.06 +/- 5.54**	**84.99 +/- 3.07**	**84.40 +/- 2.93**	**84.47 +/- 2.94**

Appendix B

Hyperparameter Tuning

Table B.1: List of the best parameters for each approach in the Random Forest models.

Parameter Tuning in Random Forest Models

Dataset	**Approach**	**Feature Selection**	**n_estimators**	**criterion**	**max_depth**	**max_features**	**min_samples_split**	**min_samples_leaf**
COAD	**Multiclass**	MRMR	100	’entropy’	3	None	6	2
		ANOVA	100	’gini’	2	None	6	2
		15-Feature Set	100	’gini’	3	None	6	4
	Binary Classification MSI-L merged with MSI-H	DR + MRMR	100	’gini’	3	None	6	2
		DR + ANOVA	100	’entropy’	3	None	6	4
		MRMR	100	’gini’	3	None	6	4
		ANOVA	100	’gini’	3	None	6	4
		15-Feature Set	100	’gini’	3	None	6	4
	Binary Classification MSI-L merged with MSS	DR + MRMR	150	’gini’	3	None	6	2
		DR + ANOVA	100	’gini’	3	None	6	2
		MRMR	150	’entropy’	3	None	6	4
		ANOVA	150	’gini’	3	None	6	2
		15-Feature Set	150	’gini’	3	None	6	4
STAD	**Multiclass**	MRMR	100	’gini’	3	None	6	2
		ANOVA	100	’gini’	3	None	6	4
		15-Feature Set	150	’gini’	3	None	6	2
	Binary Classification MSI-L merged with MSI-H	DR + MRMR	150	’gini’	3	None	6	2
		DR + ANOVA	100	’gini’	3	None	6	2
		MRMR	150	’gini’	3	None	6	4
		ANOVA	150	’entropy’	3	None	6	2
		15-Feature Set	100	’entropy’	2	None	6	2
	Binary Classification MSI-L merged with MSS	DR + MRMR	100	’gini’	2	None	6	2
		DR + ANOVA	150	’gini’	3	None	6	4
		MRMR	100	’gini’	3	None	6	4
		ANOVA	100	’gini’	3	None	6	4
		15-Feature Set	100	’gini’	3	None	6	2
UCEC	**Multiclass**	MRMR	100	’gini’	3	None	6	2
		ANOVA	150	’entropy’	3	None	6	2
		15-Feature Set	100	’gini’	3	None	6	4
	Binary Classification MSI-L merged with MSI-H	DR + MRMR	100	’entropy’	3	None	6	4
		DR + ANOVA	150	’gini’	3	None	6	4
		MRMR	100	’entropy’	3	None	6	2
		ANOVA	100	’gini’	3	None	6	4
		15-Feature Set	150	’gini’	3	None	6	4
	Binary Classification MSI-L merged with MSS	DR + MRMR	100	’gini’	3	None	6	4
		DR + ANOVA	100	’gini’	3	None	6	4
		MRMR	100	’gini’	3	None	6	2
		ANOVA	100	’gini’	3	None	6	2
		15-Feature Set	100	’gini’	3	None	6	2

Table B.2: List of the best parameters for each approach in the MLP models.

Parameter Tuning in Multi-Layer Perceptron Models

Dataset	**Approach**	**Feature Selection**	**hidden_layer_sizes**	**activation**	**alpha**	**solver**	**max_iter**	**max_fun**
COAD	**Multiclass**	MRMR	(16, 8, 16)	’identity’	0.05	’lgbfs’	200	10000
		ANOVA	(16, 16)	’tanh’	0.05	’lgbfs’	150	10000
		15-Feature Set	(16, 8, 16)	’identity’	0.05	’lgbfs’	200	8000
	Binary Classification MSI-L merged with MSI-H	DR + MRMR	(8, 8, 8)	’tanh’	0.05	’lgbfs’	200	10000
		DR + ANOVA	(8, 8, 8)	’tanh’	0.05	’lgbfs’	200	10000
		MRMR	(16, 16)	’identity’	0.05	’lgbfs’	150	10000
		ANOVA	(8, 8, 8)	’tanh’	0.05	’lgbfs’	200	10000
		15-Feature Set	(8, 8, 8)	’tanh’	0.0001	’lgbfs’	100	10000
	Binary Classification MSI-L merged with MSS	DR + MRMR	(8,)	’identity’	0.05	’lgbfs’	150	10000
		DR + ANOVA	(8, 8, 8)	’tanh’	0.05	’lgbfs’	100	10000
		MRMR	(8, 8, 8)	’identity’	0.05	’lgbfs’	200	10000
		ANOVA	(8,)	’tanh’	0.0001	’lgbfs’	100	10000
		15-Feature Set	(8,)	’identity’	0.0001	’lgbfs’	200	10000
STAD	**Multiclass**	MRMR	(8,)	’identity’	0.0001	’lgbfs’	150	10000
		ANOVA	(8,)	’identity’	0.0001	’lgbfs’	200	10000
		15-Feature Set	(8,)	’identity’	0.0001	’lgbfs’	200	10000
	Binary Classification MSI-L merged with MSI-H	DR + MRMR	(8,)	’identity’	0.0001	’lgbfs’	200	10000
		DR + ANOVA	(8, 8, 8)	’tanh’	0.05	’lgbfs’	200	10000
		MRMR	(8,)	’tanh’	0.0001	’lgbfs’	200	10000
		ANOVA	(16, 8, 16)	’tanh’	0.0001	’lgbfs’	150	10000
		15-Feature Set	(16, 16)	’identity’	0.0001	’lgbfs’	150	10000
	Binary Classification MSI-L merged with MSS	DR + MRMR	(8,)	’identity’	0.05	’lgbfs’	200	10000
		DR + ANOVA	(16, 8, 16)	’identity’	0.05	’lgbfs’	200	10000
		MRMR	(8,)	’identity’	0.05	’lgbfs’	150	10000
		ANOVA	(8,)	’identity’	0.05	’lgbfs’	200	10000
		15-Feature Set	(8,)	’identity’	0.0001	’lgbfs’	200	10000
UCEC	**Multiclass**	MRMR	(8, 8)	’identity’	0.0001	’lgbfs’	200	10000
		ANOVA	(8,)	’tanh’	0.05	’lgbfs’	200	10000
		15-Feature Set	(8, 8)	’identity’	0.0001	’lgbfs’	200	10000
	Binary Classification MSI-L merged with MSI-H	DR + MRMR	(16, 8, 16)	’identity’	0.0001	’lgbfs’	200	10000
		DR + ANOVA	(8,)	’identity’	0.0001	’lgbfs’	200	10000
		MRMR	(16, 16)	’tanh’	0.0001	’lgbfs’	200	10000
		ANOVA	(8, 8)	’tanh’	0.05	’lgbfs’	200	10000
		15-Feature Set	(8, 8)	’tanh’	0.0001	’lgbfs’	200	10000
	Binary Classification MSI-L merged with MSS	DR + MRMR	(8, 8)	’identity’	0.05	’lgbfs’	200	10000
		DR + ANOVA	(16, 8, 16)	’identity’	0.05	’lgbfs’	200	10000
		MRMR	(8, 8)	’identity’	0.05	’lgbfs’	200	10000
		ANOVA	(8, 8, 8)	’identity’	0.05	’lgbfs’	200	10000
		15-Feature Set	(8,)	’identity’	0.0001	’lgbfs’	200	10000

Table B.3: List of the best parameters for each approach in the KNN models.

Parameter Tuning in K-Nearest Neighbours Models

Dataset	**Approach**	**Feature Selection**	**n_neighbours**	**weights**	**algorithm**	**leaf_size**	**p**
COAD	**Multiclass**	MRMR	11	’uniform’	’auto’	3	1
		ANOVA	5	’uniform’	’auto’	3	1
		15-Feature Set	11	’uniform’	’auto’	3	2
	Binary Classification MSI-L merged with MSI-H	DR + MRMR	11	’uniform’	’auto’	3	2
		DR + ANOVA	11	’uniform’	’auto’	3	2
		MRMR	11	’uniform’	’auto’	3	1
		ANOVA	11	’uniform’	’auto’	3	1
		15-Feature Set	11	’uniform’	’auto’	3	2
	Binary Classification MSI-L merged with MSS	DR + MRMR	9	’uniform’	’auto’	3	1
		DR + ANOVA	11	’uniform’	’auto’	3	1
		MRMR	9	’uniform’	’auto’	3	1
		ANOVA	5	’uniform’	’auto’	3	1
		15-Feature Set	7	’uniform’	’auto’	3	1
STAD	**Multiclass**	MRMR	5	’uniform’	’auto’	3	1
		ANOVA	5	’uniform’	’auto’	3	1
		15-Feature Set	11	’uniform’	’auto’	3	1
	Binary Classification MSI-L merged with MSI-H	DR + MRMR	11	’uniform’	’auto’	3	1
		DR + ANOVA	7	’uniform’	’auto’	3	1
		MRMR	11	’uniform’	’auto’	3	1
		ANOVA	11	’uniform’	’auto’	3	1
		15-Feature Set	11	’uniform’	’auto’	3	1
	Binary Classification MSI-L merged with MSS	DR + MRMR	7	’uniform’	’auto’	3	1
		DR + ANOVA	7	’uniform’	’auto’	3	1
		MRMR	5	’uniform’	’auto’	3	1
		ANOVA	11	’uniform’	’auto’	3	1
		15-Feature Set	5	’uniform’	’auto’	3	1
UCEC	**Multiclass**	MRMR	9	’uniform’	’auto’	3	1
		ANOVA	11	’uniform’	’auto’	3	1
		15-Feature Set	11	’uniform’	’auto’	3	1
	Binary Classification MSI-L merged with MSI-H	DR + MRMR	5	’uniform’	’auto’	3	1
		DR + ANOVA	11	’uniform’	’auto’	3	1
		MRMR	11	’uniform’	’auto’	3	1
		ANOVA	11	’uniform’	’auto’	3	1
		15-Feature Set	11	’uniform’	’auto’	3	2
	Binary Classification MSI-L merged with MSS	DR + MRMR	11	’uniform’	’auto’	3	1
		DR + ANOVA	11	’uniform’	’auto’	3	1
		MRMR	9	’uniform’	’auto’	3	1
		ANOVA	5	’uniform’	’auto’	3	1
		15-Feature Set	11	’uniform’	’auto’	3	1

Appendix C

Heatmaps of the Influential Genes

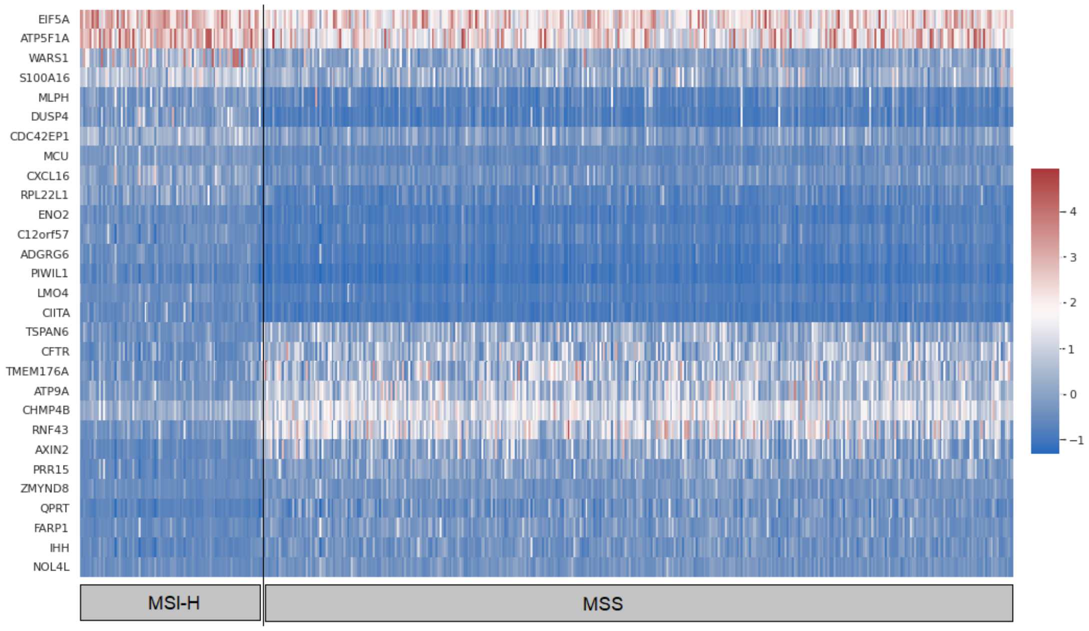

Figure C.1: Heatmap of the most influential genes selected by DR + ANOVA with the 2-class COAD dataset.

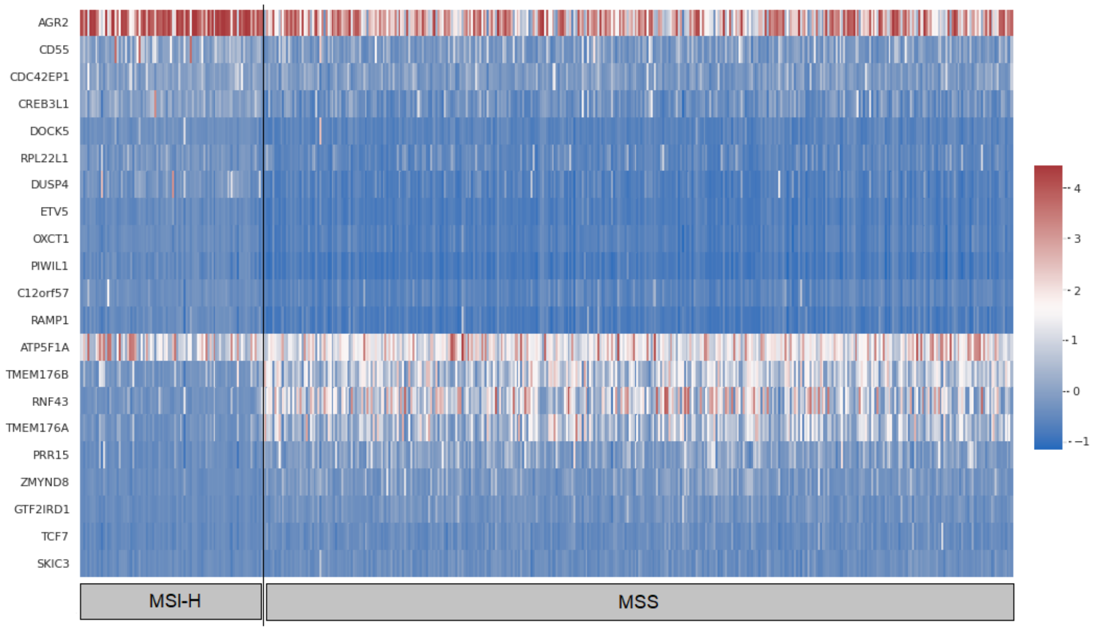

Figure C.2: Heatmap of the most influential genes selected by MRMR with the 2-class COAD dataset.

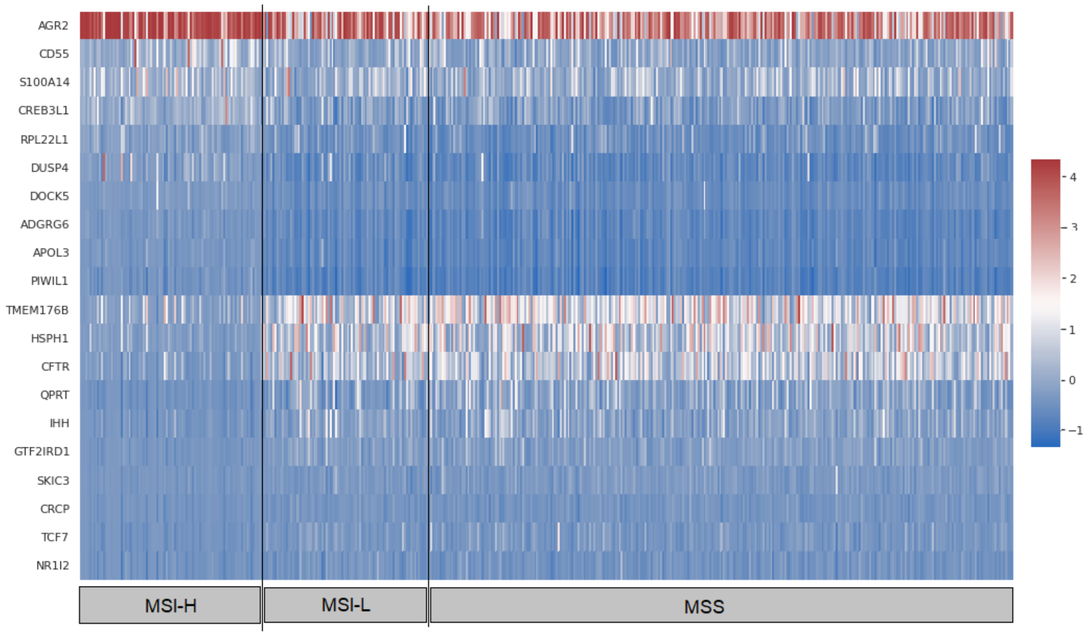

Figure C.3: Heatmap of the most influential genes selected by MRMR with the 3-class COAD dataset.

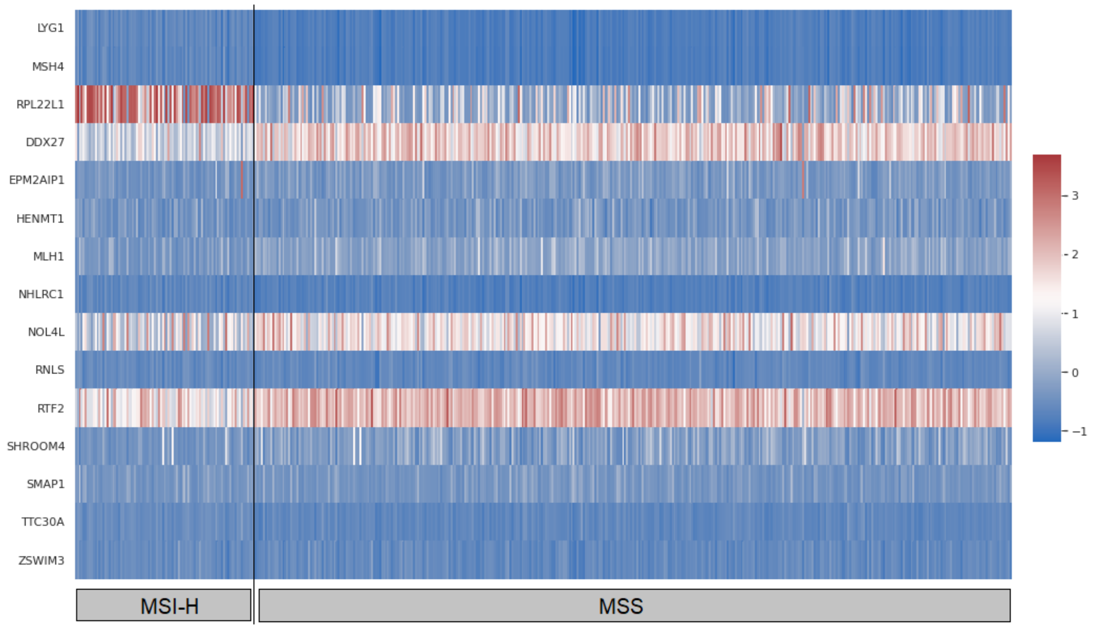

Figure C.4: Heatmap of the 15 genes selected by the 15-feature set with the 2-class COAD dataset.

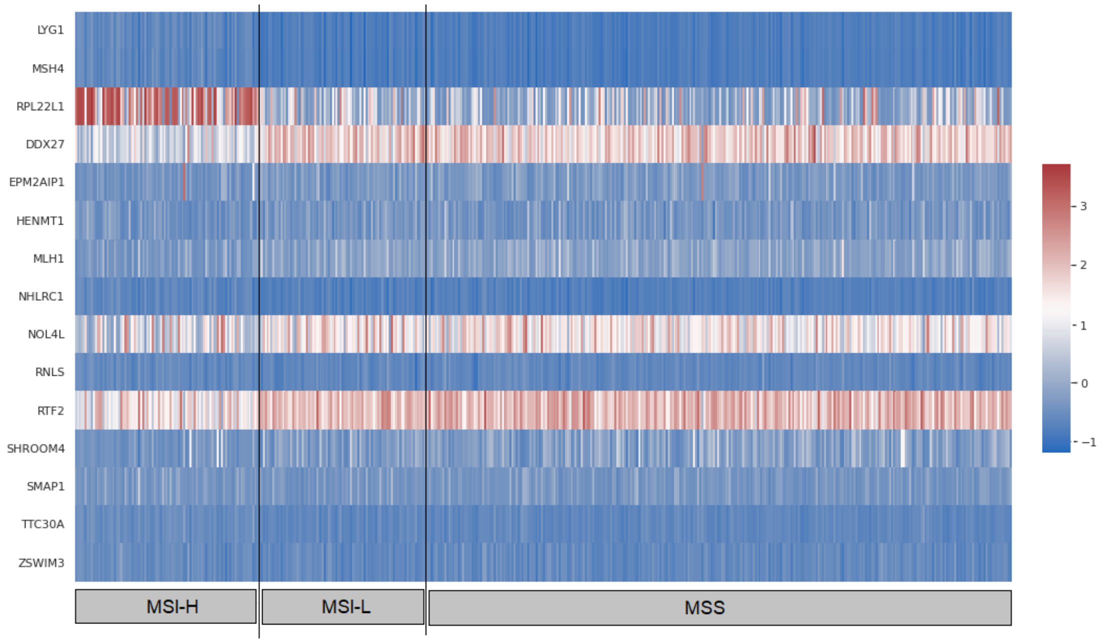

Figure C.5: Heatmap of the 15 genes selected by the 15-feature set with the 3-class COAD dataset.

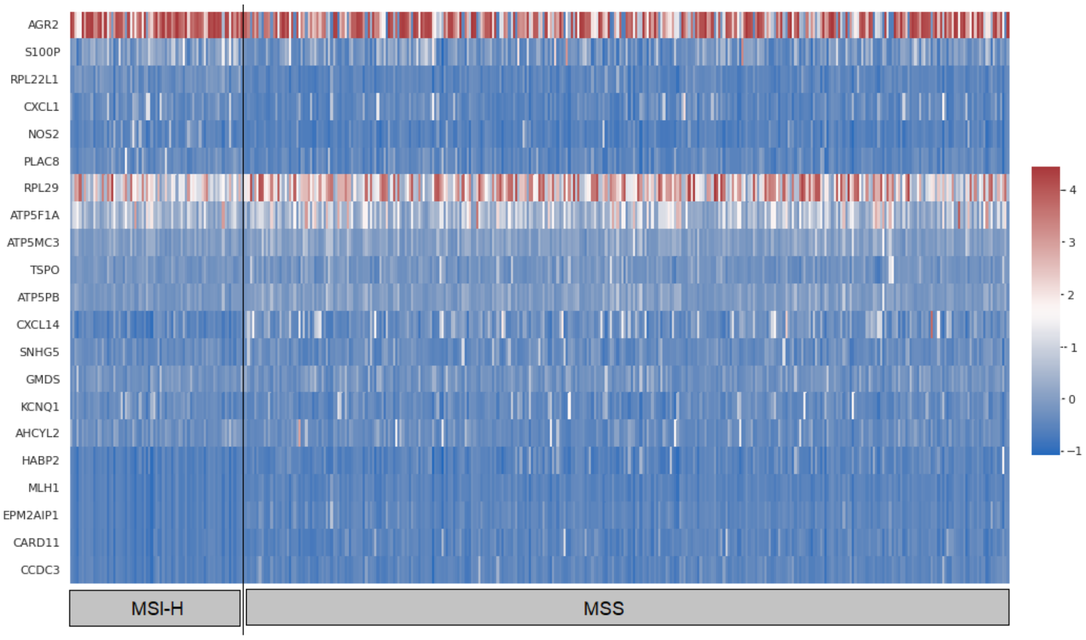

Figure C.6: Heatmap of the most influential genes selected by DR + MRMR with the 2-class STAD dataset.

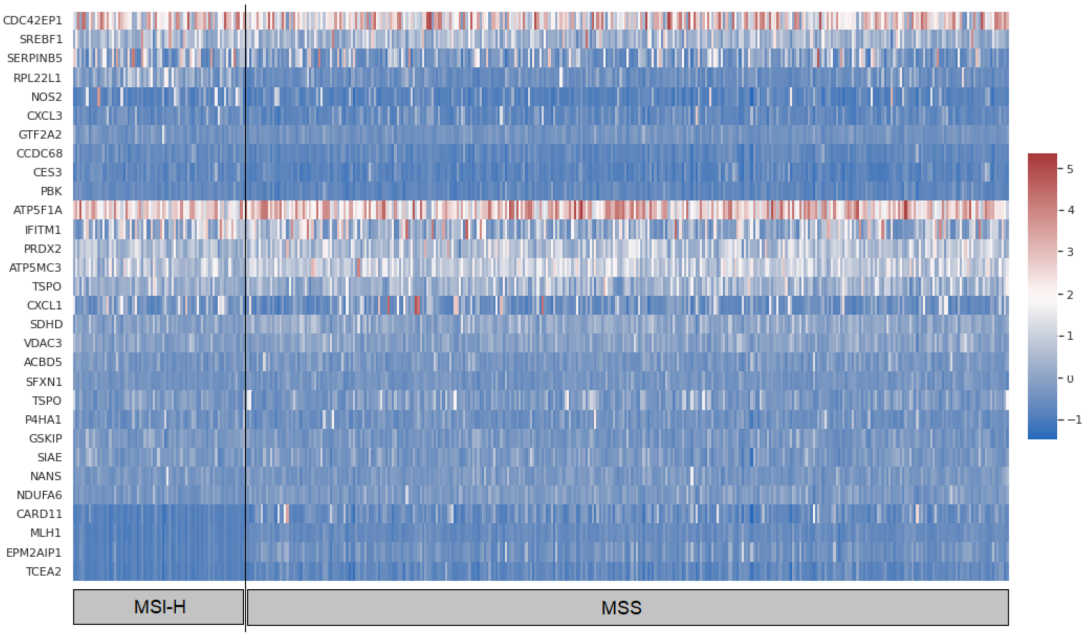

Figure C.7: Heatmap of the most influential genes selected by MRMR with the 2-class STAD dataset.

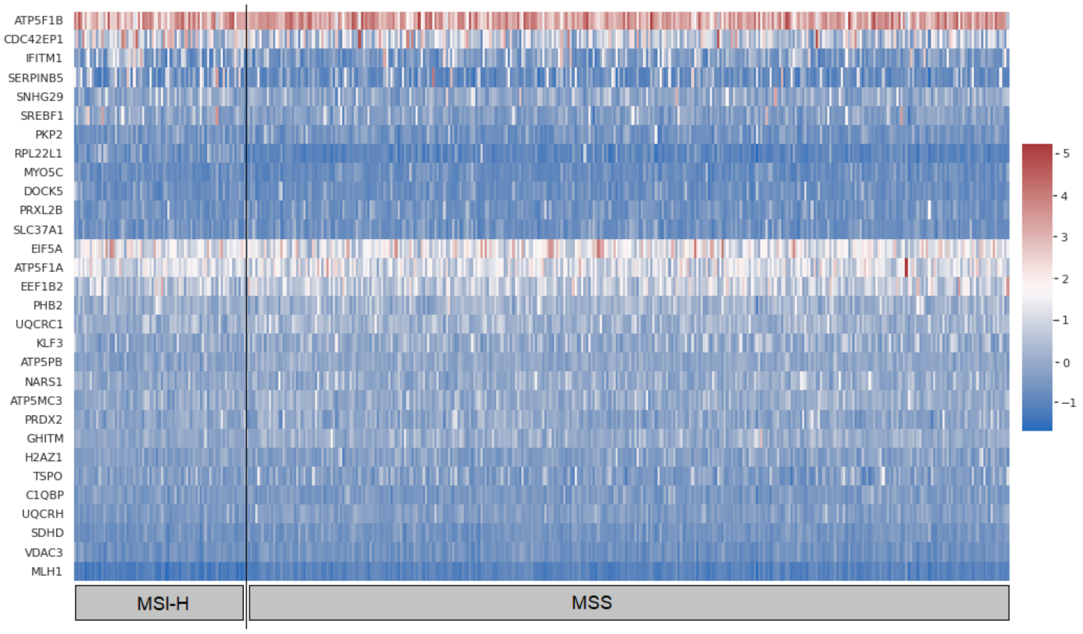

Figure C.8: Heatmap of the most influential genes selected by ANOVA with the 3-class STAD dataset.

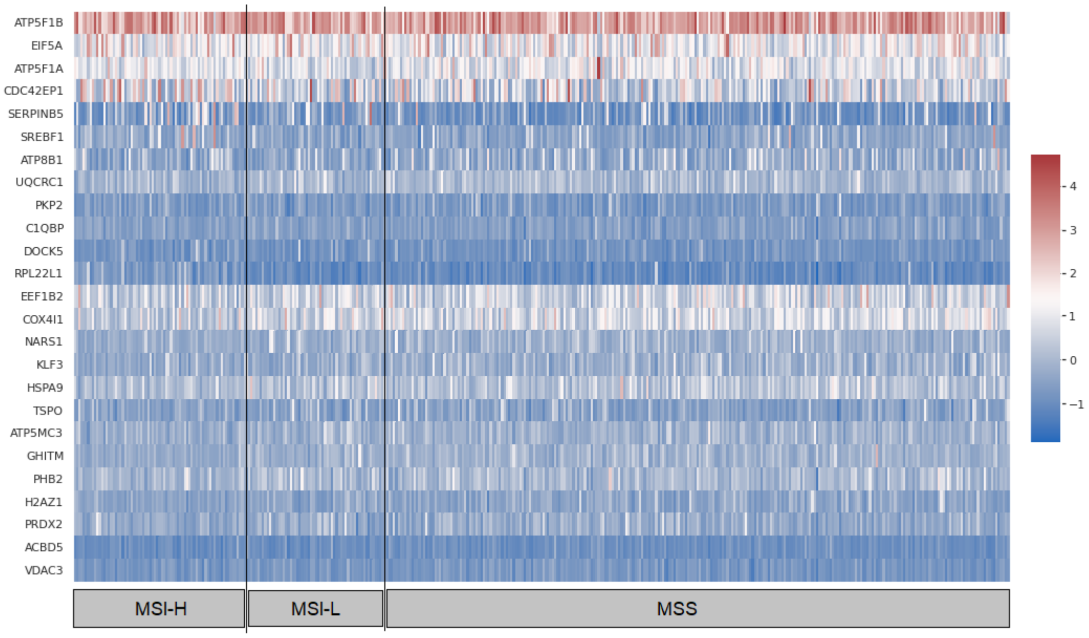

Figure C.9: Heatmap of the most influential genes selected by ANOVA with the 3-class STAD dataset.

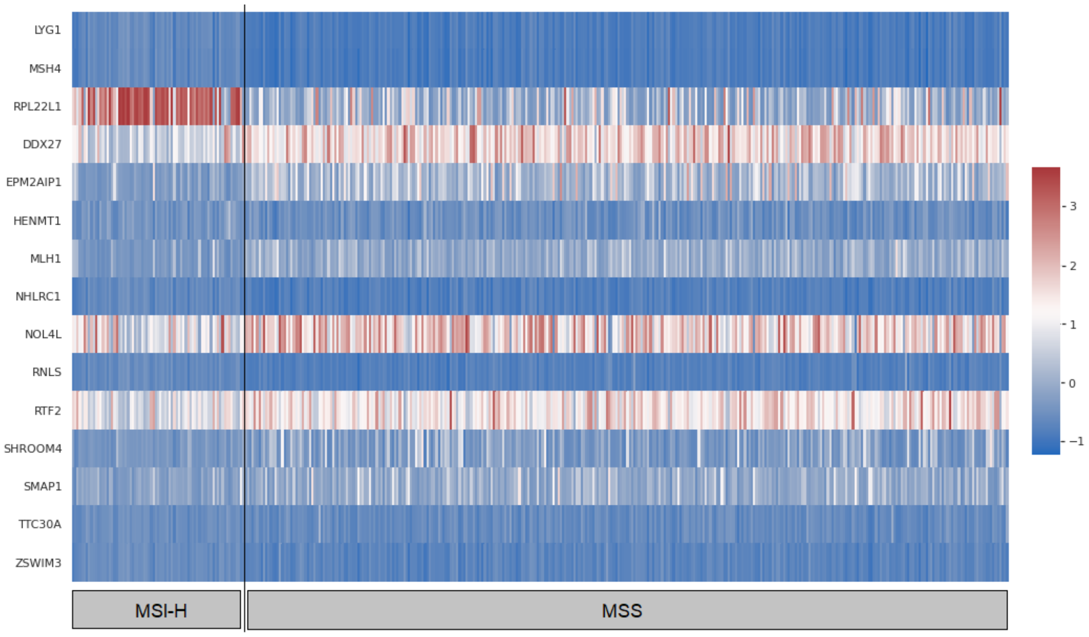

Figure C.10: Heatmap of the 15 genes selected by the 15-feature set with the 2-class STAD dataset.

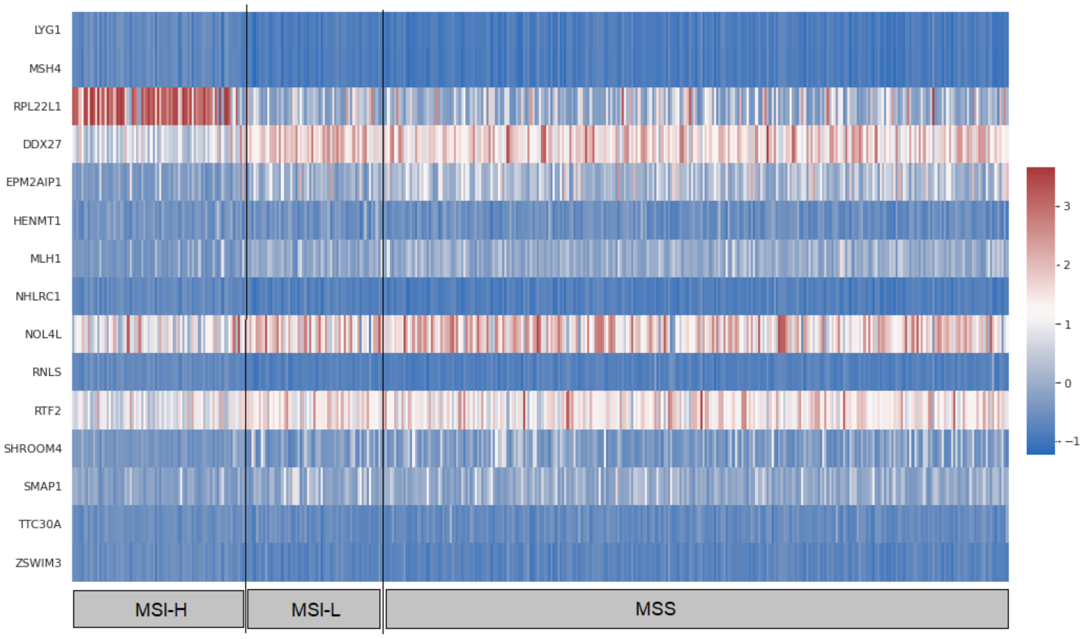

Figure C.11: Heatmap of the 15 genes selected by the 15-feature set with the 3-class STAD dataset.

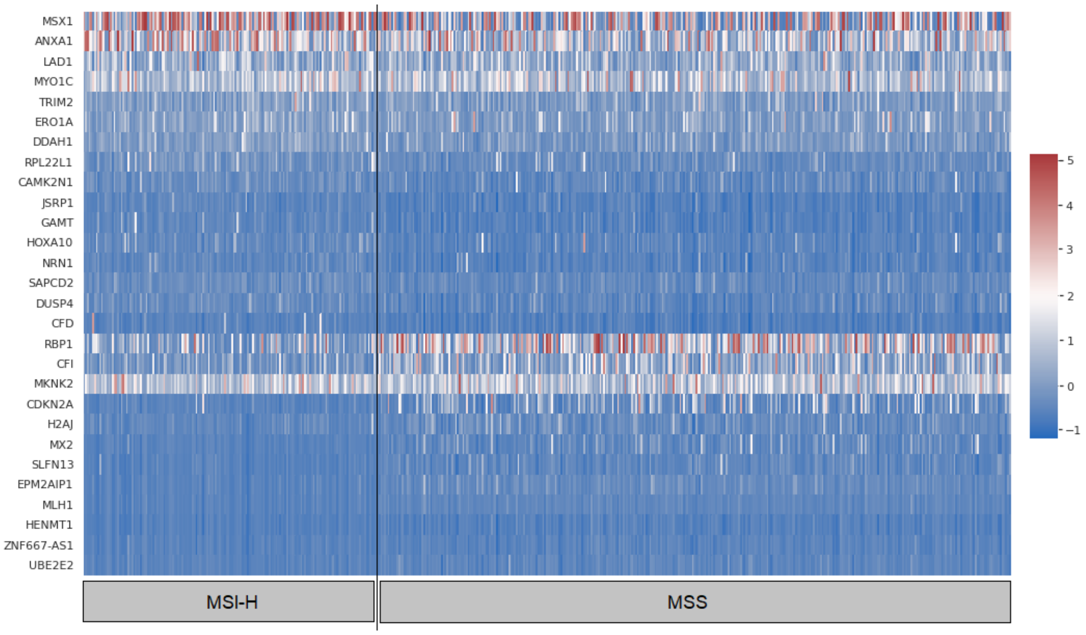

Figure C.12: Heatmap of the most influential genes selected by DR + MRMR with the 2-class UCEC dataset.

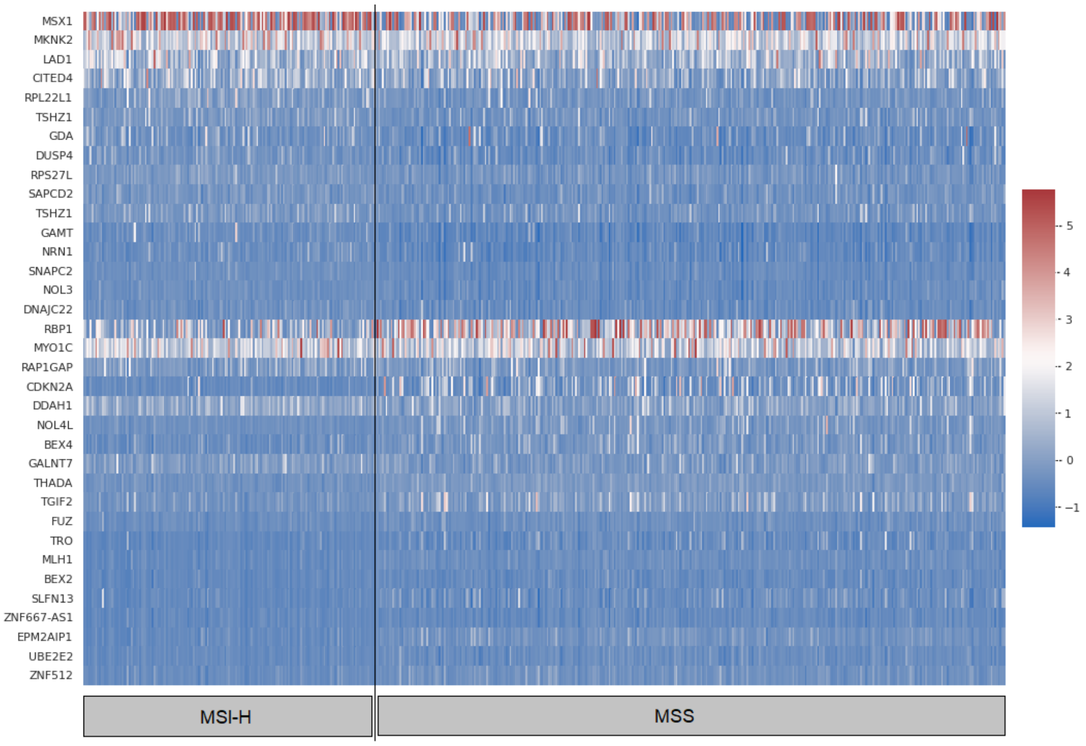

Figure C.13: Heatmap of the most influential genes selected by DR + ANOVA with the 2-class UCEC dataset.

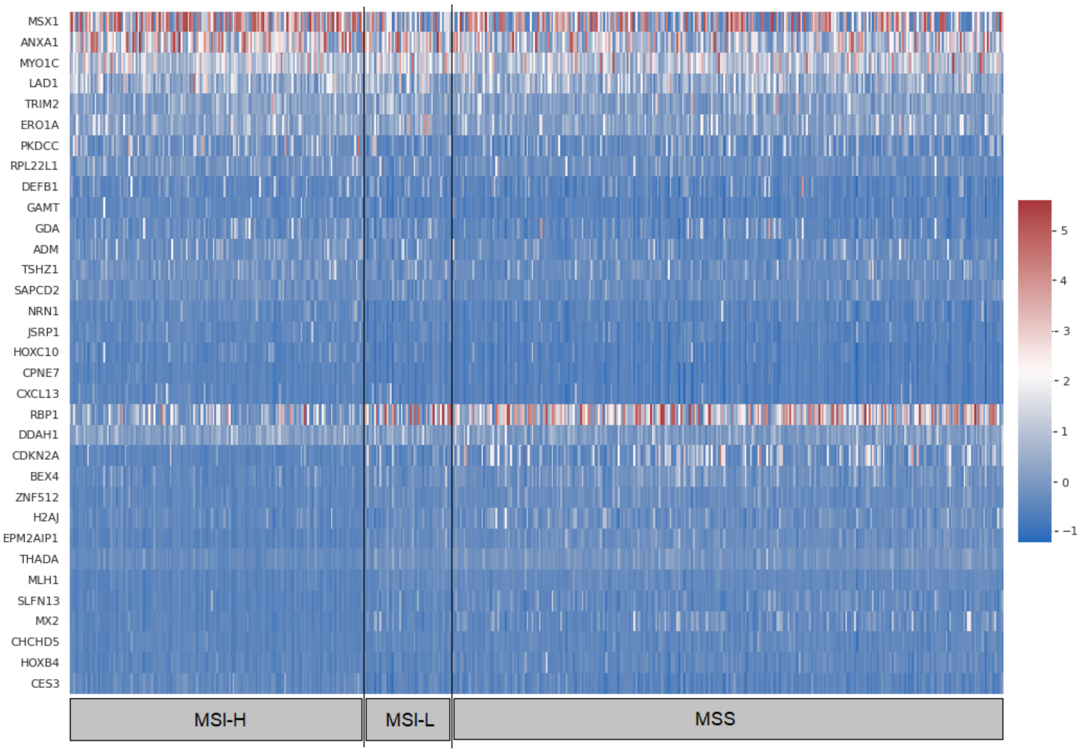

Figure C.14: Heatmap of the most influential genes selected by MRMR with the 3-class UCEC dataset.

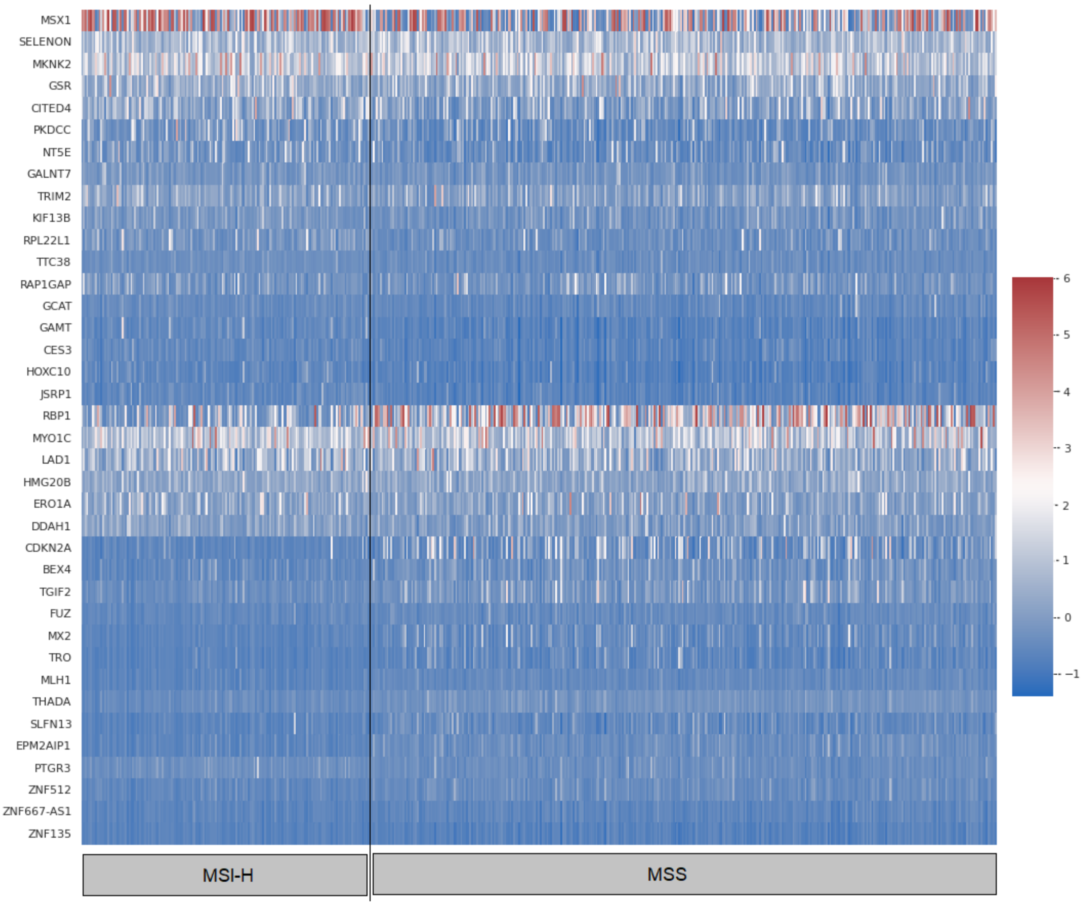

Figure C.15: Heatmap of the most influential genes selected by ANOVA with the 2-class UCEC dataset.

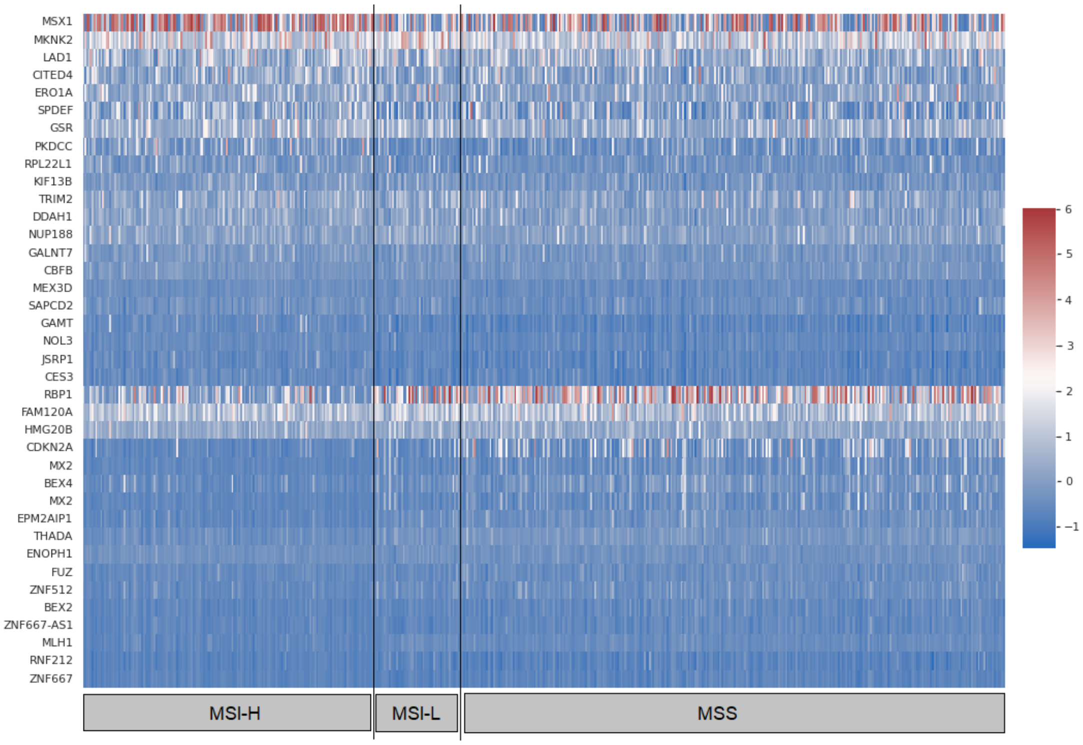

Figure C.16: Heatmap of the most influential genes selected by ANOVA with the 3-class UCEC dataset.