Dimensionality Reduction of Biomedical Tumor Profiles : a Machine Learning Approach

Martin Palazzo

Thèse
de doctorat
de l'UTT

Martin PALAZZO

Dimensionality Reduction of Biomedical Tumor Profiles: a Machine Learning Approach

Champ disciplinaire :
Sciences pour l'Ingénieur

2021TROY0031 **Année 2021**

Thèse en cotutelle avec la Universidad Technologica Nacional – CABA - Argentine

UTN.BA
UNIVERSIDAD TECNOLÓGICA NACIONAL
FACULTAD REGIONAL BUENOS AIRES

utt
UNIVERSITÉ DE TECHNOLOGIE
TROYES

THESE

pour l'obtention du grade de

DOCTEUR
de l'UNIVERSITE DE TECHNOLOGIE DE TROYES

en SCIENCES POUR L'INGENIEUR

Spécialité : OPTIMISATION ET SURETE DES SYSTEMES

présentée et soutenue par

Martin PALAZZO

le 5 octobre 2021

Dimensionality Reduction of Biomedical Tumor Profiles: a Machine Learning Approach

JURY

M. Frédéric BERTRAND	PROFESSEUR DES UNIVERSITES	Président
M. Ariel CHERNOMORETZ	INVESTIGADOR INDEPENDIENTE	Rapporteur
Mme Florence D'ALCHE BUC	PROFESSEURE TELECOM PARIS	Rapporteure
M. Emmanuel IARUSSI	INVESTIGADOR ASISTENTE	Examinateur
M. Morten NIELSEN	GRUPPELDER PROFESSOR	Examinateur

Personnalités invitées

M. Pierre BEAUSEROY	PROFESSEUR DES UNIVERSITES	Directeur de thèse
M. Patricio YANKILEVICH	INVESTIGADOR INDEPENDIENTE CONICET	Directeur de thèse
M. Hugo AIMAR	INVESTIGADOR SUPERIOR CONICET	
M. Luis MORELLI	INVESTIGADOR INDEPENDIENTE CONICET	
M. Diego TOMASSI	SENIOR DATA SCIENTIST	

Abstract

The increasing pace of data generation from tumor profiles during the last decade has enabled the development of statistical learning algorithms to explore and analyze the landscape of tumor types, subtypes and patient survival from a biomolecular point of view. Tumor data is mainly described by transcriptomic features and the level of expression of a given gene-transcript in the tumor cell, therefore these features can be used to learn statistical rules that improves the understanding about the state and type of a cancer cell.
Nevertheless transcriptomic tumor data is high dimensional and each tumor can be described by thousands of gene features making difficult to perform a machine learning task and to understand the underlying biological mechanisms. This thesis studies how to reduce dimensionality and to gain interpretability about which genes encodes signal of the data distribution by proposing dimension reduction methods based on Feature Selection and Feature Extraction pipelines. The proposed methods are based on Latent Variable Models and Kernel Methods with the idea to explore the connection between pair-wise similarity functions of tumor samples and low dimensional latent spaces that captures the inner structure of the training data. Proposed methods have shown improvements in supervised and unsupervised feature selection tasks when compared with benchmark methods to classify and learn subgroups of tumors respectively. In addition, the methods developed in this thesis have been extended to deal with dimension reduction tasks when multi-omics input data is considered.

Resumen

El ritmo creciente de generación de datos a partir de perfiles tumorales durante la última década ha permitido el desarrollo de algoritmos de aprendizaje estadístico para explorar y analizar el panorama de los tipos y subtipos de tumores y la supervivencia de los pacientes desde un punto de vista biomolecular. Los datos tumorales se describen principalmente mediante características transcriptómicas y el nivel de expresión de un determinado gen-transcrito en la célula tumoral, por lo que estas características pueden utilizarse para aprender reglas estadísticas que mejoren la comprensión sobre el estado y el tipo de una célula cancerosa. Sin embargo, los datos transcriptómicos de los tumores son altamente dimensionales y cada tumor puede ser descrito por miles de características genéticas, lo que dificulta la tarea de aprendizaje automático y la comprensión de los mecanismos biológicos subyacentes. Esta tesis estudia cómo reducir la dimensionalidad y ganar interpretabilidad sobre qué genes codifican la señal de la distribución de datos proponiendo métodos de reducción de la dimensión basados en pipelines de selección y extracción de características. Los métodos propuestos se basan en Modelos de Variables Latentes y Métodos Kernel con la idea de explorar la conexión entre las funciones de similitud por pares de las muestras tumorales y los espacios latentes de baja dimensión que capturan la estructura interna de los datos de entrenamiento. Los métodos propuestos han mostrado mejoras en las tareas de selección de características supervisadas y no supervisadas en comparación con los métodos de referencia para clasificar y aprender subgrupos de tumores respectivamente. Además, los métodos desarrollados en esta tesis se han ampliado para abordar las tareas de reducción de dimensión cuando se consideran datos de entrada multiómicos.

Résumé

Le rythme croissant de génération de données à partir de profils tumoraux au cours de la dernière décennie a permis le développement d'algorithmes d'apprentissage statistique pour explorer et analyser le paysage des types et sous-types de tumeurs et la survie des patients d'un point de vue biomoléculaire. Les données sur les tumeurs sont principalement décrites par des caractéristiques transcriptomiques et le niveau d'expression d'un gène-transcrit donné dans la cellule tumorale. Ces caractéristiques peuvent donc être utilisées pour apprendre des règles statistiques qui améliorent la compréhension de l'état et du type d'une cellule cancéreuse.
Néanmoins, les données transcriptomiques des tumeurs sont très dimensionnelles et chaque tumeur peut être décrite par des milliers de caractéristiques génétiques, ce qui rend difficile l'apprentissage automatique et la compréhension des mécanismes biologiques sous-jacents. Cette thèse étudie comment réduire la dimensionnalité et gagner en interprétabilité sur les gènes qui codent le signal de la distribution des données en proposant des méthodes de réduction de dimension basées sur des pipelines de sélection et d'extraction de caractéristiques. Les méthodes proposées sont basées sur des modèles de variables latentes et des méthodes de noyau avec l'idée d'explorer la connexion entre les fonctions de similarité par paire d'échantillons de tumeurs et les espaces latents de faible dimension qui capturent la structure interne des données d'entraînement. Les méthodes proposées ont montré des améliorations dans les tâches de sélection de caractéristiques supervisées et non supervisées par rapport aux méthodes de référence pour classer et apprendre des sous-groupes de tumeurs respectivement. En outre, les méthodes développées dans cette thèse ont été étendues pour traiter les tâches de réduction de dimension lorsque des données d'entrée multi-omiques sont considérées.

Acknowledgments

I want to thank my PhD directors Pierre and Patricio. Thanks to them I could dive deep into the disciplines of machine learning and bioinformatics. Being part of their laboratories was an amazing and inspiring experience. I really appreciate the patience and dedication you have had to guide me and help me develop my research path.
A special mention to Raul Sack and Fernando Gache, who have invested effort and time from the Argentinian side to develop the co-tutelle PhD program between UTN and UTT. We remember Raul as a great leader and mentor.
My friends from UTN Rama, Sofi and Nico were amazing phd colleagues from whom I have been able to learn a lot, they have also been an excellent company in the difficult moments of the doctorate.
At UTT I could meet valuable people like Samir, Nacef, Aleja, Andrea, David and Guillaume. Thanks to them my days at UTT have been the best.
I am grateful for having spent 4 years at the Ibioba institute. There I was not only able to learn biology and work alongside the best scientists in the country, I was also able to meet wonderful people like Fiore, Ivan, Diego, Sol, Gabi, Ludmi, Marina, Flor, Dani, Jero, Belu, Sebastian, and David.
To my dear Mora Matassi, who has given me the company and strength I needed to submit this thesis. Merci chérie.
Also thank my lifelong friends, who have supported me and been available whenever I need their help and energy: Ivan, Diego, Tedy, Agus, el Mono, Esteban, Juancho liqui, AleStaro, Emilia, Tomas, Amit, Meli Breda, Meli Lopez, Franco, Anita, Chona, Nano, Nati cristo, Loli, Pri, Choique, Javi, Juancho JT, Sol, Santi and JC.
To the colleagues and friends I have made thanks to academia and think-tanks: Victoria Peterson, Valentin Muro, Enzo Ferrante, Carlos Sarraute, Rodrigo Maranzana, Paloma Urtizberea, Diana Mosquera.
Being able to discuss science with Diego Tomassi and Javier Burroni was a pleasure and a very important learning stage for me.
I want to thank to my friends and family from France: Virgile, Eric, Meri, Monica, Anais, Jean Louis and Jean Claude. They made my life in France a great experience.
Finally, thanks to my family, who have supported me unconditionally in all the stages and challenges that I have faced, it is a pride for me to be part of you: Marcos, Sandra, Andres, Sofia, Tomas, Marco, Grace, Anibal, Elvira, Amado and Loraine.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

The bioinformatics community is facing an important increase of biomedical data thanks to the next generation sequencing technologies. Tumors from cancer patients can be sequenced and recorded from a molecular point of view at high pace. In bioinformatics the bottleneck is not anymore the data generation process but the data interpretation and processing. A biomedical tumor profile can be linked to the clinical attributes of the patient such as survival, tumor type and subtype. Nevertheless the biomedical tumor data is described by thousands of biomarkers making difficult the interpretation and analysis. For this reason the large availability of tumor data is an opportunity to learn statistical rules from the data to improve eventually the diagnosis and treatment and a challenge due to its high dimensionality.
Machine learning makes possible to learn hidden rules from large volumes of data. The majority of real case applications of machine learning pipelines use high dimensional data such the biomedical case. Reducing dimensionality is a necessary step to process and interpret the input data. To reduce dimensionality two approaches are studied in this thesis: feature selection and feature extraction. Feature selection aims to select a subset of the input dimensions to form a low dimensional space. Feature extraction aims to project the high dimensional input data in a low dimensional space using all the input dimensions. Despite both methods are useful to reduce dimensionality the feature selection one offers interpretability of the reduction since the selected features are kept from the input data. On the other side dimension reduction via feature extraction may capture better the latent structure of the input data using all the input dimensions but without specifying which one contribute the most to the data distribution.
To improve supervised and unsupervised machine learning tasks on biomedical tumor profiles this thesis proposes to couple feature selection and feature extraction methods. The aim of this thesis is to propose supervised and unsupervised feature selection and feature extraction pipelines applied on gene expression data from tumor profiles with the objective to improve tumor classification and tumor clustering respectively.
The target audience of this thesis are researchers, engineers and practitioners in machine

learning and bioinformatics who are interested in dimension reduction of biomedical tumor profiles from humans cells.

1.1.1 Outline of this thesis

The chapter 2 of the thesis studied how data of tumor profiles is structured, why it is high dimensional and the challenges and opportunities of learning from this data.

The third chapter studied the supervised feature selection problem for tumor classification. In this chapter a novel regularization term is proposed named *Kernel Latent Regularization* by selecting features that explain the tumor labels and the latent structure of the input data simultaneously.

The chapter 4 extends the Kernel Latent Regularization approach of the previous chapter in feature extraction, more specifically in Sufficient Dimension Reduction problems.

The chapter 5 studies the unsupervised feature selection problem to select gene signatures that groups the tumor profiles by learning a latent space with feature extraction methods.

On chapter 6 the multi-omic data integration problem is studied and used to improve unsupervised feature selection by learning a latent space from multi-omic data.

Finally chapter 7 details the final conclusions of the thesis.

All the scripts and codes used in this thesis are available from the following github repository `https://github.com/martinepalazzo/phd_thesis_biomed_dim_red`.

Chapter 2

Biomolecular data of tumor profiles

2.1 Introduction

The first chapter of this thesis entitled *Dimensionality reduction on biomedical tumor profiles: a machine learning approach* introduces the challenges and opportunities that the machine learning and pattern recognition community has for the exploration of cancer disease in a biomedical context. These opportunities are mainly propelled by the continuous growth and availability of large biomedical and molecular data from tumor profiles such as DNA, RNA, Proteins and Metabolites [1]. Additionally from the mathematical and statistical point of view machine learning methods have been improving their capabilities to process high dimensional and complex data during the last decades. Finally, the improvement of computational processors allow the implementation of the learning methods in large datasets efficiently [2].
This thesis chapter is composed by five sections. Section 1 is the current introduction. Section 2 defines Cancer Disease and why is important the research and development of methods to better understand the complexity of tumors. In section 3 tumor data is detailed and it explains how molecular measurements from tumor profiles are stored in large datasets covering different modalities known as *omics*. The section 4 defines and formalizes clinical data from cancer patients by explaining how tumor profiles can be categorized among primary sites, subtypes, stages and survival. Section 5 formalizes the machine learning approaches and how these can be used to take advantage of the large availability of tumor data with the objective to improve cancer diagnosis and biomaker discovery in biomedical research. Finally section 6 summarizes the challenges and opportunities that the machine learning community has in cancer genomics research from a biomedical perspective.

2.2 Cancer disease

Cancer is a group of genetic diseases that can be developed in any location of the human body [3]. It originates when old or damage cells that should die instead survive and grow

uncontrollably altering the normal function of a cell. Then these cells start dividing without stopping forming a cluster of cells known as tumor. Cancer disease occurs as a result of the changes and mutations produced within the DNA sequence in the genome of the cancer cells [4]. The worst scenario is named *metastasis* and it is when the growth of tumor cells of a given organ continues and in consequence forcing these to spread into different sites of the body making multiple organs fail and eventually cause death.
The *World Health Organization* (WHO) estimates in 2018 that 9.6 million people die which is one every six deaths caused by cancer positioning it as the second largest cause of death globally. Figure 2.1 shows the different primary sites of tumors in the human body.

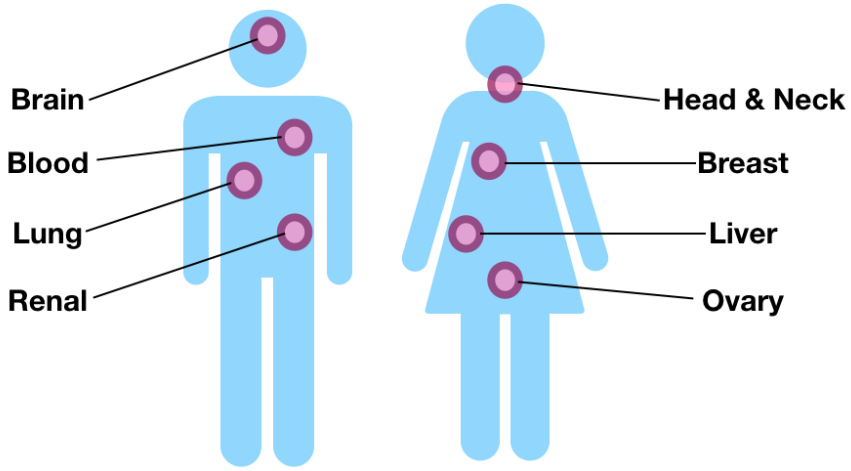

Figure 2.1: Primary sites of tumors

Despite each tumor can be characterized by the same set of biological features like DNA, RNA, Proteins or Metabolites, the signal encoded within these will differ significantly depending on the primary site in the body where the tumor is located, the stage of the tumor and the subtype. We name this encoded signal as *genetic signature* referring to how active different regions and parts of the human genome are within specific tumor cell giving each tumor a specific identity and state. The term *encoded signal* in this thesis is reffered to the regions of the cell genome and transcriptome that help to differentiate or find similarities between tumor profiles.
Therefore if d biological features are measured these can be modeled as random variables $X_1, X_2, ..., X_i, ..., X_d$ that take different values across different tumor samples ruled by a probability distribution $p_1(x_1), p_2(x_2), ..., p_i(x_i), ..., p_d(x_d)$. For example, a tumor located in the Brain will present a different genetic signature than a tumor located in the Liver. Even when tumors belong to the same type and primary site like Kidney (Renal) these can be categorized into different subtypes such as Clear Cell Carcinoma (KIRC) or Papillary Carcinoma (KIRP) where each one presents a different genetic signature associated to each

subtype. In addition, the genetic signature of a tumor of the same type and subtype may differ according to the *prognosis* of the patient, those with low or high survival rate.
It is possible also that the encoded signal of a given genetic signature is located in just a fraction of the human genome. This situation reveals the need to search the region or area of the genome responsible of encoding the signal in order to understand which biological mechanisms are involved. The genetic signature concept is formalized in section 3 which explains the biomedical tumor data.
For the reasons described above the processing and interpretation of the genetic signatures of tumors is a crucial task to understand the complexity of cancer and to identify accurately a given tumor by type, subtype, stage or prognosis. It is important to remark that unlike preventing cancer before it appears as a disease, this thesis focuses on accurately characterizing a tumor once it already exists in the patient. Determining the type, subtype, stage or associated prognosis of a tumor from a genetic signature allows to select the appropriate treatment to be applied to each patient in a context of personalized medicine. Furthermore, it is possible to discover bio-markers that facilitate the interpretation of the genomic data of the tumor in order to improve the decision-making process in cancer diagnosis.

2.3 Tumor omic data

Each tumor is characterized by different groups of molecular features. Each of these groups is a layer of features characterized by the same types of macro-molecule known as *omics*. When a tumor is described by different omics simultaneously then the context is called *multi-omics* [5]. The multi-omics context described fits within the concept of multi-modal data, a situation commonly used when features are obtained from different sources and environments such as images and natural language for example [6]. The multi-omics context of cancer involves groups of molecular features such as DNA, RNA, Proteins or Metabolites, thus each tumor can be described by multiple omics from biomolecular data.
Additionally each tumor has a phenotype which is a term associated to the observable characteristics and traits. The phenotype of a tumor is caused and determined by both the omic features and the environment. Since cancer disease has a strong genetic and multi-omic component [7] in this thesis the focus and study is on tumor omics and how these molecular features are related to a given phenotype. Tumor phenotypes can be classified by the primary site where the tumor has first developed and by histological type [8] which is associated to the type of tissue of the tumor. Generally the primary site defines the type of tumor while the tissue or histology defines the subtype of tumor. Tumor type and subtype define the phenotype which corresponds to a clinical diagnosis made by clinical doctors. Moreover, tumor data is labeled by phenotype but also by clinical information like tumor stage, patient survival, patient age and gender.
Each omic contains potential biomarkers that could help to detect a tumor outcome like tumor subtype, stage or patient survival. For this reason biomarker discovery is an important task to identify key biomedical features. Moreover, biomarker discovery allow medical doctors to focus only in a reduced and small set of features instead of measuring the full genome of a

patient.

2.3.1 Omic data

Nowadays the advance of biotechnology allow the collection of almost any type of molecular data from a living organism. There are different types of macro-molecules which can be measured by different *omic* technologies like Next Generation Sequencing [9] and RNA sequencing [10]. Then a tumor can be characterized by multiple omic data. This availability of multi-modal data allow the exploration and analysis of tumors from different perspectives but at the same time a new problem arises: how to deal with heterogeneous and high dimensional sources of biological data. Each omic layer can present tens of thousands of features like genes or mutations which means from a machine learning perspective a high dimensional representation space. There is a need to learn simpler and low dimensional representations of these complex systems for supervised and unsupervised statistical learning approaches. The analysis of a set of tumor samples characterized by single or multiple omics makes possibly to learn statistical rules between the genomic signature and the phenotype such as the subtype, the stage or even the survival rate of the patient known as prognosis.
Omic data is ruled by a specific flow of information determined by the Central Dogma of Biology. The dogma defines how genetic information flows through different omics, from DNA to messenger RNA and to protein in order to determine the function of a cell and finally the phenotype of the organism [11] inmersed in a given ambient and context. Figure 2.2 details the multi-omics structure of a tumor cell. Each cell has its DNA sequence composed by four types of nucleic acid bases grouped in genes. A tumor cell presents Somatic Mutations within genes which corresponds to single changes in one of the bases of the DNA sequence forming a somatic mutation signature for each tumor. Then the mutated DNA sequence and the genetic material of a gene is copied and *expressed* as RNA-messenger molecules known as *Transcripts* within the cell forming an expression profile. A molecular machinery transforms the expressed Transcripts into Proteins which are molecules that have assigned a specific function based on their structure. Therefore the genes are fragments of the DNA chain that encode a specific function for a given cell and serves as building bricks to form a protein. Since not all the genes are expressed and activated at the same time with the same intensity the protein landscape within each cell will be different as well as its function. Finally the protein landscape will determine the phenotype of the tumor. Each tumor will present a specific molecular signature within each omic. These signatures may help to predict and estimate the phenotype such as the tumor subtype or tumor stage.
In this thesis a single omic is mainly used, processed and analyzed by machine learning models: Gene Expression or Transcriptomics. Additionally in the last chapter of the thesis Transcriptomics is combined also with Somatic Mutations or Genomics as a multi-omic problem. Machine Learning models are applied on these omics to perform biomarker discovery via feature selection and to perform supervised and unsupervised tasks like classification or clustering of tumor subtypes.
This section presents the two omics used in this thesis. To clarify the Gene Expression or

Transcriptomics explanation the Somatic Mutation omic or Genomics has to be introduced first.

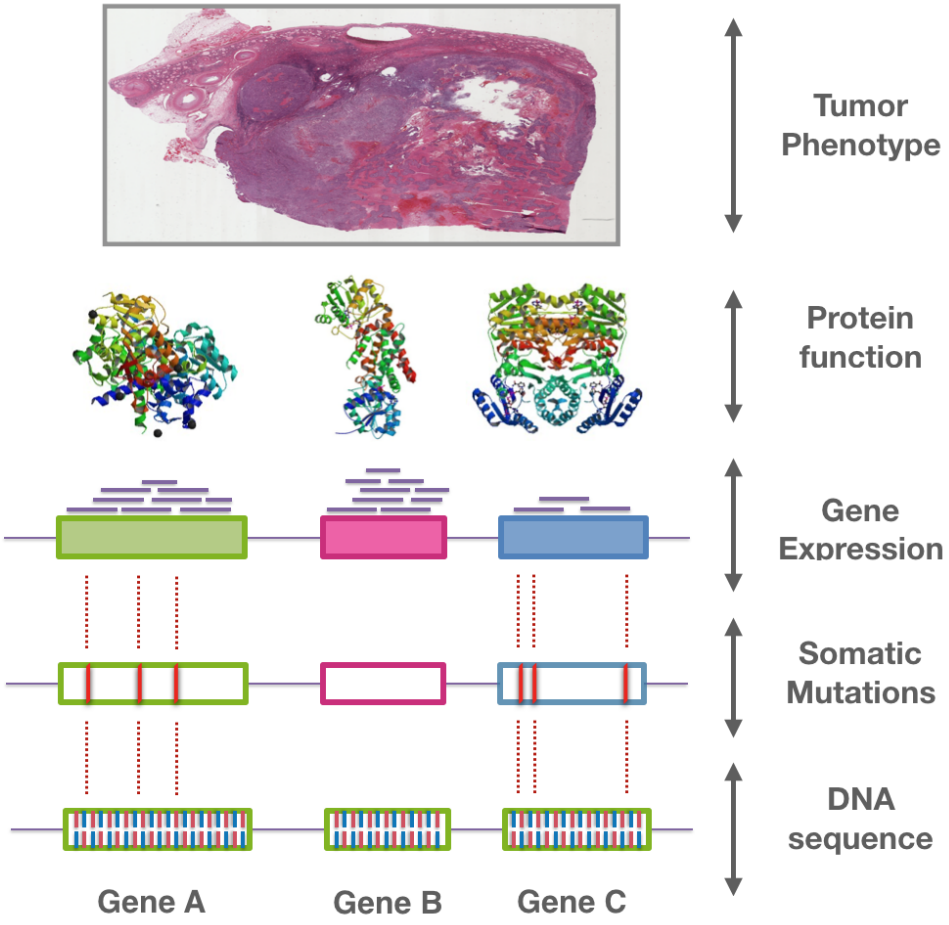

Figure 2.2: Diagram of multi-omic tumor data

2.3.2 Somatic Mutations

The DNA is a long sequence composed only of four bases: Ademine (A), Timine (T), Citosine (C), Guanine (G). Expression 2.1 gives an example of an extract of a DNA sequence

$$\text{AGTCAGCGGTATCT} \tag{2.1}$$

where the full DNA sequence forms a much longer chain of bases. One of the genomic layers used in this thesis is composed of Simple Somatic Mutation (SSM) data [12]. The SSM data is the one that characterizes the set of somatic point mutations of tumor. A simple somatic mutation is an alteration of one base in the DNA that occurs during the life of the organism after conception. SSM can be found in any cell of the human body and are not transferred

to the offspring. These alterations can be a deletion, an insertion or a substitution of a base to another one.

$$\begin{array}{rcccc} \text{Reference sequence :} & \text{AGTCAGC} & \mathbf{G} & \text{GTCACT} \\ \text{Mutated sequence :} & \text{AGTCAGC} & \mathbf{C} & \text{GTCACT} \end{array} \tag{2.2}$$

Expression 2.2 shows an example of a mutation of one single base from G to C. These somatic mutations affect a particular gene and eventually downstream the gene expression and protein function. The somatic mutation found on each gene can be considered a random variable that takes different values across different tumor profiles thus these build a specific signature for each tumor. The somatic mutation signatures can be used to characterize the phenotype of a tumor. Therefore SSM can be used as biomarkers for patient diagnosis. Table 1.1 details how the Somatic Mutation data is available

Field	Values
Donor ID	DO228283
Project Code	BRCA-US
Chromosome	3
Chromosome Start	27763452
Mutation Type	single base substitution
Mutated from Allele	C
Mutated to Allele	T
Consequence Type	Exon Variant
Gene affected	ENSG00000163508

Table 2.1: Annotations from a Simple Somatic Mutation.

Each mutation belongs to a patient whose cancer has been diagnosed by a tumor type and subtype. This tumor is located in a primary site of the body of the patient and is recorded in the context of a Project Code. The somatic mutation is located at a position within a chromosome, is characterized by the type of mutation, is labeled by its effect which can affect a protein coding gene. In that specific case the corresponding gene is identified. A protein coding gene is the region of the genome that contains the necessary information to build a protein with a specific function in the organism. Therefore a somatic mutation that occurs within a protein coding gene region is a potential cause of a protein malfunction. The genome of a cancer cell can present thousands of simple somatic mutations. Some of these mutations will occur outside of critical regions and are known as *introns* and *intergenic* areas. Other mutations occur within critical regions like *exons* where there is a direct impact on the protein aminoacids, the building blocks of proteins. Nevertheless analyzing each somatic mutation separately is a difficult challenge due to the large number of mutations involved simultaneously. In addition, somatic mutations can interact between each other and generate a compound effect on the phenotype of the tumor.
To deal with the complexity of a compound effect of thousands of somatic point mutation

$$X_{mut} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1d} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{nd} \end{bmatrix}$$

Figure 2.3: Somatic Mutation Data matrix.

features across hundred of cancer patients a multi-variate analysis done using machine learning and pattern recognition methods is developed during this thesis. To process the data a $n \times d$ samples by features matrix X is needed. This matrix has as many rows as patients and as many columns as mutated genes. In this thesis, when somatic mutation data is used, the biological features are the protein coding genes that contains somatic mutations.

As detailed in Figure 2.3 each row represents a patient and each column a mutated gene. The value x_{ij} of patient i in gene j is equal to the number of somatic mutations within gene j. A characteristic of Simple Somatic Mutation data is the sparsity of the matrix x since there are almost 20.000 protein genes but each patient also has mutations in a few of these. The sparsity is an issue when using machine learning approaches since there are not enough mutated genes in total to describe a density across all the sample set for a specific gene.

2.3.3 Gene Expression

As mentioned in the previous section the mutational and histological features can be used to estimate and determine the phenotype of a tumor. Nevertheless these features are not informative enough in some cases and thus other biomedical features are needed [13][14]. Then a new layer of biomedical information that characterizes tumor profiles arises with a better correlation with clinical features: gene expression or *Transcriptomics* [15]. Gene expression corresponds to the concentration level of RNA messenger molecules associated to each gene in a cell. Almost every cell in an organism contains the same exact copy of DNA, nevertheless each cell expresses with different intensities each gene through RNA messenger defining an expression profile and thus a cell function. Therefore, the relative expression between genes across multiple tumor cell samples can bring valuable information.
Table 2.2 shows how the raw gene expression data is obtained. Each patient or donor is registered by the Project Code which is associated directly to a tumor type and subtype. For one tumor, all the protein coding genes are listed. Each gene is described by the Chromosome where it is located, the gene ID and the Read Count which is the number of RNA transcript measured. The Read count is a measure of how intensive is the expression of a gene and serves for relative comparison of expression between genes. Then for a patient i and a gene j the expression level is represented as x_{ij} and it is defined as the count of read fragments of RNA associated to gene j.

$$X_{exp} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1d} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{nd} \end{bmatrix}$$

Figure 2.4: Gene Expression Data matrix.

Field	Values
Donor ID	DO228283
Project Code	BRCA-US
Chromosome	3
Gene ID	ALPK2
Gene Ensemble ID	ENSG00000163508
Read Count	278
Normalized Read Count	1.3467

Table 2.2: Annotations from the expression of a Gene.

The pre-processing step for gene expression data consists in building from the data of Table 2.2 a $n \times d$ matrix X_{exp} as detailed in Figure 2.4. The rows represent tumor patients and columns protein coding genes similar to the Figure 2.3

The main difference between Somatic Mutation and Gene Expression matrices relies in the type of values each matrix represent. The Somatic Mutation matrix contains integer numbers representing the number of mutations within a gene j given a patient i while the expression matrix contains the level of expression as a real number of a gene j given a patient i. The first one is a sparse matrix since the somatic mutations only occur in a small fraction of genes and the majority of the positions of the matrix are zero. The second one contains values in almost all the positions.

2.4 Clinical data

The International Cancer Genome Consortium (ICGC) is the source clinical and biological data used in this thesis [16]. Each multi-omic tumor profile is associated with multiple clinical labels.

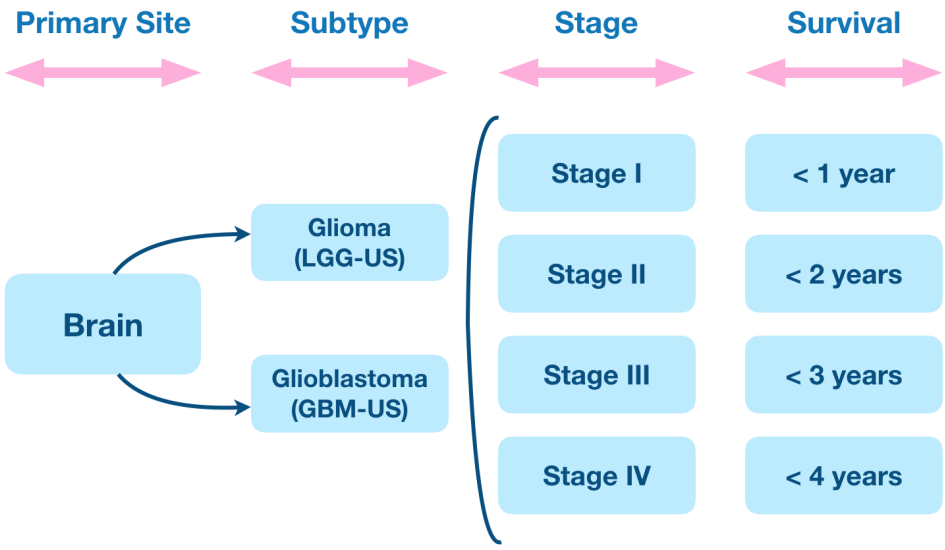

Figure 2.5: Tumor clinical labels.

These labels are Primary Site, Tumor Subtype, Tumor Stage and Survival days since diagnosis and are detailed in Figure2.5 4 with an example on brain cancer.
Primary site or tumor type indicates the organ where the tumor appears within the human body. Within a primary site tumors can present different characteristics and properties thus a sub categorization is possible with the tumor subtype labels. The set of tumors labeled with the same subtype are grouped together in a same Project ID associated to the organization that has been responsible to publish the omic and clinical data of the tumor. Projects are defined by tumor subtype and country. Then each tumor profile can be categorized by stage if it has been diagnosed when the tumor started to grow or by late stage if it has been diagnosed when it has a significant size. Finally, some tumor profiles are labeled with the survival days from diagnosis.
Table 3 shows an example of the tumor clinical information where each row corresponds to

Donor ID	Project ID	Primary Site	Age	Gender	Stage	Survival days	Vital status
DO232761	BRCA-US	Breast	63	Female	NaN	1539	Alive
DO222843	MELA-AU	Skin	76	Male	II	907	Deceased
DO50970	LICA-FR	Liver	72	Female	NaN	NaN	Alive
DO6480	CLLE-ES	Blood	60	Female	A	8579	Deceased
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
DO48854	NBL-US	Nervous System	4	Male	IV	232	Deceased

Table 2.3: Clinical information of tumor patients.

a patient or donor and each column details different clinical labels.
The ICGC data portal has published omic and concerning clinical information from multiple tumor types. This consortium centralizes 86 cancer projects conforming more than 15 tumor primary sites. Each cancer project is associated to a specific tumor subtype given a tumor type and country. In total there are more than 10.000 donors who provided tumor samples. Each tumor sample is characterized at least by one of the omic layers described in the previous section and in some cases tumor profiles are described by multiple omics.
One of the hypothesis considered in this thesis is that molecular omic data (gene expression, somatic mutations and others) can be used as predictors to estimate a clinical outcome like Tumor Subtype, Tumor Stage or Surivival days. These clinical outcomes are assumed to be dependant of the omic data [17]. For this reasons the following section formalizes the importance to find a statistical dependance that explains the clinical outcome of a patient by just observing omic data via machine learning models.

2.5 Tumor data for machine learning tasks

In the previous sections the molecular omic data and the clinical data have been presented and described. In this section the aim is to explain how machine learning models can be used in this context to estimate clinical outcomes from omic data.
Let an input space $\mathcal{X}$ be a d-dimensional space such that $\mathcal{X} \subseteq \mathbb{R}^d$. Then a set S_u of n samples is defined in $\mathcal{X}$ as

$$S_u = \{x_1, x_2, ..., x_n\}$$

where each sample x_i is a d-dimensional vector. In this thesis each sample vector is associated to a tumor cell profile characterized by a specific omic data. Each omic dataset is composed of n tumor profiles characterized by d gene features as shown in the last section.
If the clinical information of table 2.3 is available then a label vector y can be defined as $y \subseteq \mathbb{R}$ for continuous clinical variables like survival days, $y = \{-1, 1\}$ for binary categorical labels like tumor subtype or $y = \{1, 2, ..., T\}$ for multiple categorical labels like primary site or tumor stage where T is the number of clinical categorical labels. The set of samples S_u can be re-defined as a set S_s of input-output pairs of tumor samples and clinical labels

$$S_s = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$$

The set S_u corresponds to an unsupervised set of samples since there is not any clinical label paired with the samples while the set S_s is considered as a supervised one since each sample is paired with a clinical label. In this thesis I propose both scenarios when tumor clinical labels are available and those where labels are not available named supervised and unsupervised problems respectively [18]. Figure 2.6 shows an example cartoon of a tumor dataset composed of somatic mutations and gene expression labeled with clinical information from patients.

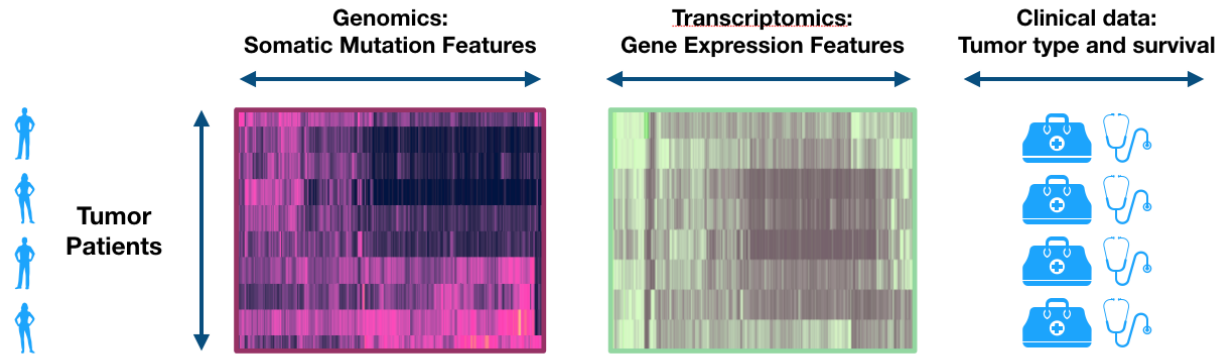

Figure 2.6: Diagram of multi-omic and clinical data

In the case of a supervised approach the goal is to learn from the set S_s a function $f(x)$ that maps a tumor sample x to a tumor label y such as

$$f(x) \sim y \tag{2.3}$$

by minimizing a loss function $L(y, f(x))$ that measures the discrepancy between the output of $f(x)$ and the true labels y. f is trained by optimizing the loss function using the dataset S_s. When the labels are continuous $y \subseteq \mathbb{R}$ the problem is defined as regression. When the labels are categorical or discrete $y = \{-1, 1\}$ the problem is defined as classification.
In the cases of unsupervised contexts where tumor labels y are not available then the kind of problems to solve are related to learning the similarities between tumor samples from S_u and finding communities or sub-groups of tumors that are highly similar between each other. The resulting groups are named *clusters* and is expected a correlation between these and clinical labels. Figure 7 shows an example of supervised and unsupervised learning problems.

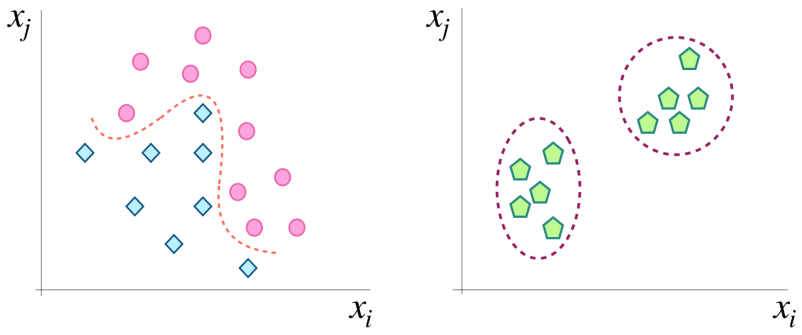

Figure 2.7: Left figure: a supervised classification problem. Right figure: an unsupervised clustering problem.

Additionally, since each tumor profile is characterized by more than 20.000 protein coding genes the space where the samples lies is high dimensional. Thus an important machine learning task to be studied in this thesis is *dimensionality reduction*. The goal of dimensionality reduction is to map and characterize the input data in a lower dimension. For a machine learning algorithm high dimensional spaces and low sample sizes increase the complexity of the problem so that learning from data in this context is less efficient [19]. To measure the relation between sample size and dimensionality the sample to feature ratio r is computed as

$$r = \frac{n}{d} \tag{2.4}$$

where $r > 1$ means a number of samples higher than the number of features and $r < 1$ means a number of samples lower than the number of features. The first case is desirable since there are enough samples to describe the space $\mathcal{X}$ while the second case is not desiderable since there are not enough samples to describe the high dimensional space. The omic datasets of tumors used in this thesis show in the majority of cases a situation where $r << 1$ and the number of features far exceed the number of samples. This situation leads to ill posed training problems and is related to the state that the number of samples needed to reach a given estimation quality increases exponentially with the dimension. This situation is known as *The Curse of Dimensionality* and is not desiderable for statistical learning. In addition, it is assumed that the high dimensional input space contains several noisy features that can be discarded thus the intrinsic dimensionality of the data is assumed to be lower than the input one [20]. Figure 2.8 shows a toy example of how the sample to feature ratio varies before and after dimension reduction.

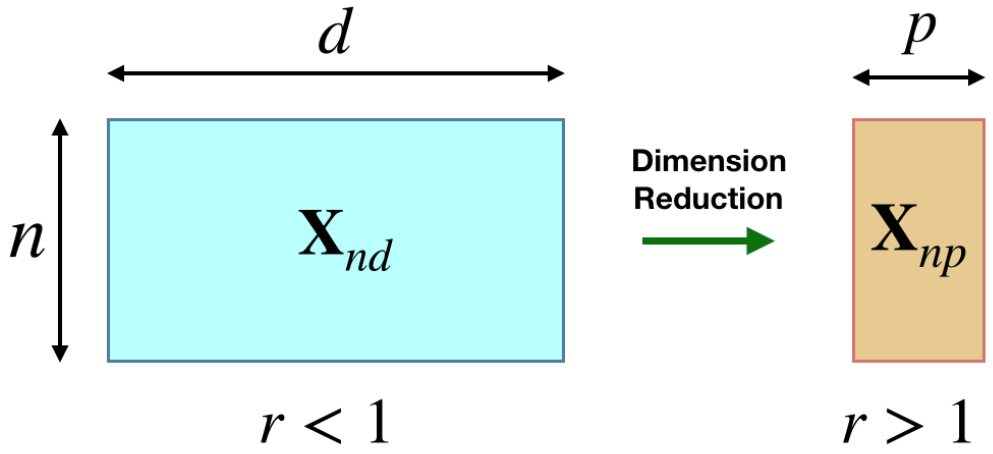

Figure 2.8: Left: Small sample to feature ratio and high dimensional space. Right: High sample to feature ratio and low dimensional space after dimensionality reduction.

For this reason learning low dimensional representations of the input data and reducing the dimensionality is an important task studied in this thesis and there are two important hypothesis to consider: firstly it is not necessary to learn from the full set of biological features and a reduced subset of genes is just enough to perform supervised or unsupervised machine learning task. Secondly there exists simpler and low dimensional latent representations of the tumor profiles that helps to understand the underlying distribution and structure of the data.

2.5.1 Exploratory Data Analysis

This subsection is devoted to explore the omic data of cancer profiles and obtain a first overview of how the data is composed. As a case example the Breast Cancer samples from the BRCA-US project of the ICGC data portal is used. Table 2.4 shows the number of tumor samples and gene features of the dataset.

Primary Site	Project Name	Gene Features	Tumor Samples
Breast	BRCA-US	20385	1041

Table 2.4: BRCA-US data.

It is observed that sample to feature ratio is $r = 0.05$ which show the low sample size in comparison with the input space dimension. Similar values of sample to feature ratio are observed in all the tumor datasets in ICGC.

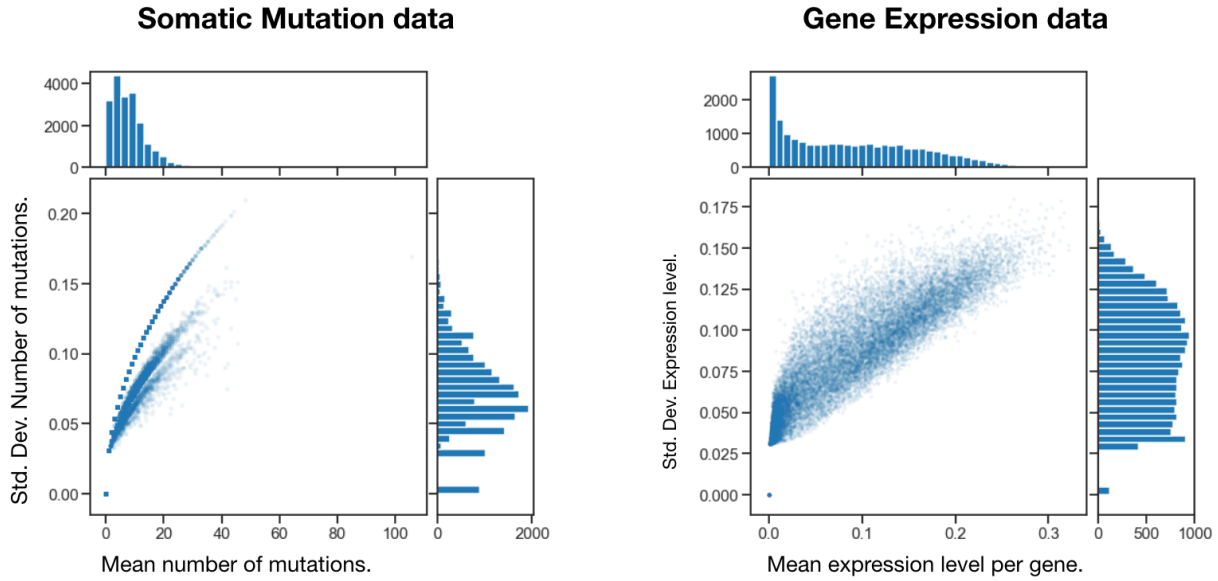

Figure 2.9: Mean and standard deviation distribution of each of the 20385 gene features values across all patients in BRCA-US dataset.

Figure 2.9 plot the mean value against the standard deviation of each gene feature for both Somatic Mutation and Gene Expression omic datasets of the BRCA-US project. A visual positive correlation appears between the mean value and the standard deviation of each feature. It is observed that the gene features that are more mutated or more expressed presents a higher standard deviation across tumor profiles. On the opposite the gene features with lower mutation numbers or expression levels present a lower variation across tumor profiles. There is a subset of gene features that presents 0 standard deviation and 0 mean mutations or expression at the bottom left of each subfigures.
Figure 2.10 shows a histogram to understand the distribution on the sum of mutations and expression of each gene feature across all patients for both omic datasets. The figure reveals that the Gene expression data has more gene features with non zero elements across all patients while the somatic mutation data presents a high number of null features and a number of genes with non-zero values in a small number of patients. Figure 2.10 shows the sparse nature of the somatic mutation data since the data matrix is filled mostly with zero elements while the gene expression data has non-zero elements in almost all the positions of the data matrix.

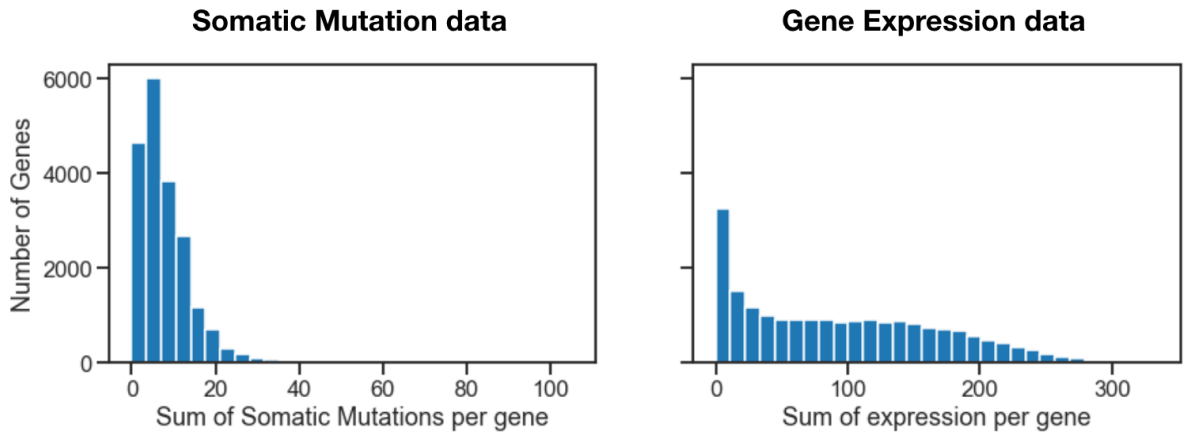

Figure 2.10: Sum of the values of the gene features across all patients of dataset BRCA-US.

Both figure 9 and figure 10 suggest that the useful information from omic data for machine learning tasks is encoded in a reduced subset of gene features thus there are genes that can be discarded.

2.6 Challenges to be addressed

The omic and clinical data of tumors presented invite to study and develop machine learning models to assess different problems. Clinical information like tumor subtype, tumor stage or survival days can be estimated from omic data using a supervised approach. In addition,

unsupervised approaches are also needed since not always the clinical labels are available thus discovering tumor clusters that present high similarity between tumor profiles is possible and of valuable interest.
Either for supervised and unsupervised approaches, the omic data used is high dimensional and the number of tumor samples low. This high dimensional scenario makes it difficult to perform supervised or unsupervised tasks. Figure 6 shows that there are gene features more informative than others in both somatic mutation and gene expression datasets. For this reason developing robust dimensionality reduction methods is necessary to extract useful information from omic data. Dimensionality reduction can be performed by selecting a subset of genes p where $p << d$ or to learning a new representation of low dimensionality by combining the input features and filtering the input noise.
Finally, figure 9 and 10 shows that gene expression data has more features with higher variation across tumor profiles than the somatic mutation data thus gene expression data is assumed to be more informative and supporting that gene expression data has been widely used for cancer classification [21]. For this reason this thesis will use gene expression data only in single omic datasets while in the last chapter of this thesis multi-omic data integration is used to combine somatic mutation and gene expression data.

Chapter 3

Supervised Feature Selection with Kernel Methods

3.1 Introduction

The *omic* tumor profiles landscape is composed and categorized by a wide range of different clinical classes such as tumor types, subtypes and patient survival making tumor profile classification a key task for machine learning in cancer informatics. Tumor classification implies to find a mathematical decision function used as a decision boundary between tumor classes in a high dimensional space $\mathcal{X}$ defined by the input gene expression features which are considered as random variables. The core idea of classification is to learn the decision function in a way that it assigns accurately clinical labels to tumor profiles of unknown clinical context. The tumor profile classification problem using gene expression data is high dimensional since it is characterized by tens of thousands of gene expression features as introduced in the chapter 01. Moreover, generally cancer datasets have a low sample size. As mentioned in the previous chapter the sample size n to feature space dimension d ratio of these datasets is $r << 1$ which characterizes situations where the curse of dimensionality is an issue of first importance. To tackle a tumor classification problem in a high dimensional context reducing the dimensionality of the data is an important task to be done. Feature selection is a family of dimensionality reduction methods designed to discard the noisy and non useful features while retaining only a subset of the input features useful to perform a supervised or unsupervised machine learning task [22]. It is assumed that the selected features contain relevant signal for a supervised learning task such as classification while the discarded ones contain noise or redundant signal. Another important goal of feature selection methods is to guide the biomarker discovery process [23] [24]. Since tumors are characterized by tens of thousands of features taking measurements of all the biological features is expensive and inefficient. Moreover, selecting a subset of gene features may improve a supervised learning task like tumor classification or survival regression and also improves a biomedical interpretation since selected features provides a way to understand the underlying biological system. Feature Selection has been widely applied in bioinformatics and cancer genomics using gene expression

data on supervised learning problems [25] [26] [27]. The hypothesis sustained in this work states that a reduced subset of genes is necessary to perform and improve significantly the supervised learning task, particularly in this work the supervised task is tumor classification. The classification task in cancer genomics takes as input a given gene expression tumor profile and assign it a phenotype label by implementing algorithms that take decisions wrt the optimization of a loss function. These algorithms learns statistical rules from labeled data in order to build a decision function.
Popular supervised learning methods such as Support Vector Machines and Perceptron or unsupervised ones such as Principal Component Analysis are using an inner product $\langle x_i, x_j \rangle$ between data samples during the learning process which can be interpreted as a similarity measure between sample x_i and sample x_j [28]. Despite this measurement allows the usage of a family of linear functions they can be restricted and limited in real-world applications like high dimensional omic tumor data. To extend the capability of those methods the introduction of kernel methods enable to use these linear framework to deal with more versatile non linear functions. Their study are at the core of this thesis [18]. Kernels have been widely used in computational biology for supervised tasks like classification or regression and for unsupervised tasks like clustering as well as to solve dimensionality reduction problems [29] [30]. Thus one can expect that Kernel methods can also improve the learning task on omic data.

In this context a set of questions arises: given a gene expression profile how can we classify tumor samples based on clinical attributes? If the data is high dimensional how can be reduced to a low dimensional space? If linear methods may be limited in high dimensional tumor data how can nonlinear ones be used to improve the learning task? In this chapter supervised learning is used in tumor classification given a gene expression profile. Feature selection is used to select a subset of genes that improves the classification performance and that helps to interpret the biological signature involved in the classification task. Kernel Methods are used to perform and improve both tumor classification, feature selection methods and are base of a novel proposed approach for feature selection tasks on gene expression data. This chapter first introduces in the following order classification algorithms, kernel methods and feature selection models. The classification section defines linear classifiers, the support vector machine classification method and the evaluation metrics in classification. The kernel methods section introduces the kernel trick and how kernels can be used to deal with classification problems where classes are not linearly separable, the Kernel Alignment metric and Multiple Kernel Learning. Then the following section is about the feature selection approach, benchmark methods are presented and performance metrics are defined. Finally a new approach named *Kernel Latent Regularization Feature Selection*(KLR-FS) [31] is proposed and experimental results are detailed.

3.2 Supervised Learning

In chapter 01 the proposed hypothesis is based in the idea that clinical outcomes and tumor phenotypes can be estimated from omic biomolecular data of tumors, particularly gene expression. Clinical variables like tumor type, subtype or survival are called *responses*, *dependent variables* or *labels* and are represented by an output label y. It is assumed that gene expression data from tumor profiles is sampled from a set of high dimensional feature space $\mathcal{X} \in \mathbb{R}^d$. To train the learning algorithms a dataset is obtained from a set N of n realizations of the random variable X with $i = 1...n$ and x_i is the i^{th} sample of the dataset. Each input vector x_i is associated to a tumor sample and each dimension j of the input vector x_{ij} is a gene expression feature. More conceptually, one can consider a set S of n labeled samples

$$S = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\} \tag{3.1}$$

that can be splited in two subsets S_{tr} as *training* and S_{te} as *test*. The training set is used by a supervised learning algorithm to learn a function $f_w(x) = \hat{y}$ parametrized by w to be used as the inference rule of labels y given x [32]. When y represents two categories or classes, labels are typically $y \in \{-1, 1\}$ and the supervised problem is named *binary classification*. A classification problem consists in finding a decision rule $f_w(x)$ that defines a boundary between the samples of the two possible classes $y \in \{-1, 1\}$.

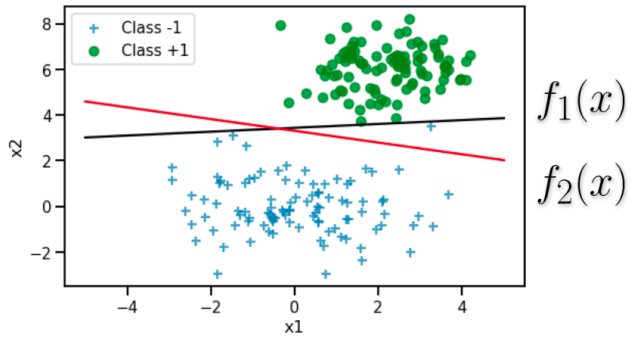

Figure 3.1: A scatter plot of two synthetic classes with $n = 200$ and $d = 2$. Each class is sampled from a two dimensional gaussian density. Black and Red boundaries are examples of two possible classification functions f_1 and f_2 respectively.

Figure 3.1 shows a two class example with $X \in \mathbb{R}^2$ with two proposed functions $F = \{f_1, f_2\}$ used as decision boundary. The selection of a final function will depend on the performance of each one in the classification task determined by how similar is the predicted output label $\hat{y}$ to the ground truth y. To determine the discrepancy between $\hat{y}$ and y a *Loss Function* $\mathcal{L}$ defined as

$$\mathcal{L}(y, \hat{y}) = \mathcal{L}(y, f_w(x)) \tag{3.2}$$

is used to penalize every time $\hat{y}$ is different from y. To find the parameter w^* of the decision function $f_w(x)$ an optimization problem is presented with the objective to minimize the Loss Function as

$$w^* = \underset{w}{\operatorname{argmin}}\ \mathbb{E}\left(\mathcal{L}(y, f_w(x))\right) \tag{3.3}$$

where the expected value of the loss can be expressed as $J(f_w)$

$$w^* = \underset{w}{\operatorname{argmin}}\ J(f_w) \tag{3.4}$$

It is desired that the learned function $f_w(x)$ predicts labels $\hat{y}$ similar to y, to achieve that goal the expectation of the Loss function is estimated from the training data S_{tr} by the empirical loss estimator J defined as

$$\widehat{J}(f_w) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(y_i, f_w(x_i)) \tag{3.5}$$

which is the mean loss computed between ground truth labels y_i and the model prediction $f_w(x_i) = \hat{y}_i$.
In a classification problem the decision functions belongs to a parametrized family or class of functions F. Training a decision function means picking one function of the class F or choosing the parameter w that characterize it so that $f_w(x)$ minimizes the empirical Loss $\widehat{J}(\hat{f}_w)$ when using a finite training dataset S_{tr} let call the learned function $\hat{f}_w$. The learned $\hat{f}_w$ may not necessary be the best decision function that can be obtained from F due to estimation error related to the finite size n of the training data. Ideally, the best decision rule obtained by the minimization of the empirical loss on F can be expressed as $\inf_{(f\in F)}J(f_w(x))$. Thus the difference between the error from the learned decision function and the optimal one from F is named Estimation Error and is expressed as

$$J_{estim} = \widehat{J}(f_w) - \inf_{(f\in F)}J(f_w(x)) \tag{3.6}$$

Additionally it is possible that the optimal decision function that minimizes the classification loss of a given problem lies outside F, let call it f^*. Therefore the difference between the optimal Error within F and the one obtained from the optimal decision function f^* is named Approximation Error and is defined as

$$J_{approx} = \inf_{(f\in F)}J(f_w(x)) - J(f^*) \tag{3.7}$$

Finally, the objective of the learning approach is to find a decision function $\hat{f}_w$ that minimizes the sum of the Estimation and the Approximation error

$$J_{mod} = \left(J(\hat{f}_w) - \inf_{(f\in F)}J(f_w(x))\right) + \left(\inf_{(f\in F)}J(f_w(x)) - J(f^*)\right) \tag{3.8}$$

known as Modeling Error J_{mod}. Figure 3.2 shows a visualization of how these errors are related in the supervised learning problem

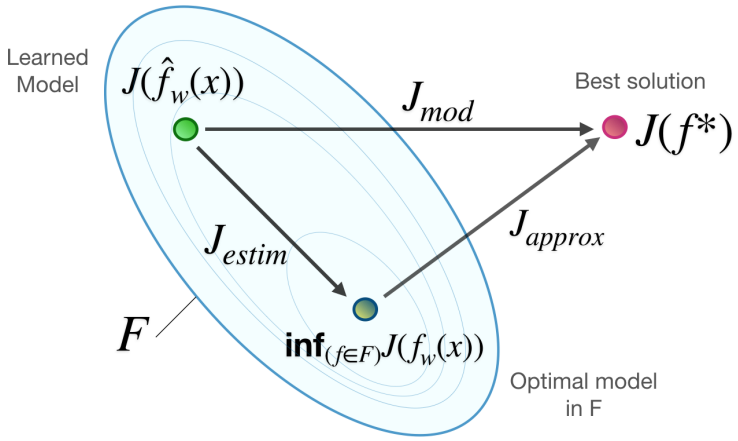

Figure 3.2: A diagram of the estimation error, approximation error and modeling error.

When F is a small set then J_{estim} is going to be limited but J_{approx} may be large. On the contrary if a very large set F is considered J_{estim} may be large and J_{approx} limited. For this reason there is a need to find a compromise between the diversity of the functions allowed by F and the estimation error.
The same idea can be reformulated from the model complexity perspective. The error of the learned function $\widehat{J}(\hat{f}_w)$ can be decomposed in two main components: The Variance and the Bias [33].

$$\widehat{J}(y, \hat{y}) = \text{Var}(\hat{f}(x)) + \left(\text{Bias}(\hat{f}(x)\right)^2 \tag{3.9}$$

Equation 3.9 suggests that a good supervised learning model may achieve a nice compromise between Bias and Variance.
The *Variance* is associated to the amount by which the estimation of the function $\hat{f}$ will change if it is trained with a different training set sampled from the same probability density function which also stands for the estimation error. If a supervised learning model has high variance then the $\hat{f}$ will change considerably given small changes in the training set. A desired situation is when a learning model estimates $\hat{f}$ with small changes when training data varies. If the estimated $\hat{f}$ is more complex than the inherent structure of the data then it is expected to obtain high variance. By this way over-complexity is penalized by the variance error.
The *Bias* term is associated to the approximation error that is considered when approximating real-life datasets. For example, when a biological system which is almost sure that is generated by a non-linear function is modeled by a linear one then it is expected to have high bias. Then the bias increases when the estimated model $\hat{f}$ is not complex enough to explain the true function that generates the data f^*.
The Estimation error can be interpreted as a consequence of the Variance while the Approximation error is a consequence of the Bias. In a ideal case the induction principle is convergent when the number of samples $n \to \infty$ then the function set F can be enlarge enough so that the variance and bias tend to 0.

These sources of error are bounded by the data availability N and the size of the function space F.
Regarding the generalization capacity of the learned model in regard to new unseen test data, a more complex and flexible method may present a higher variance and lower bias relation in test set, thus the variance error increases the overall error. On the opposite case, when the model is less complex and less flexible the difference in error between train and test set will not be high although not low enough to achieve a good classification performance.
As described in the previous sections the bias error tend to be related to the structure and distribution of the training data and the variance to the size of the available function set F. These insights fuels the idea to propose in this work a supervised learning method presented in section 5 that considers both the data structure via latent variable models and multiple kernel functions reducing in consequence the bias and variance error correspondingly. Then these improvements are proposed to reduce the classification error of tumor samples given their gene expression profiles.

3.2.1 Classification with hyperplanes

One of the objectives of this thesis is to perform tumor classification by type, subtype, stage and survival among others. The classification problems studied in this thesis are binary which means tumor labels $y \in \{-1, 1\}$ defined the two classes. To perform classification a decision function $\hat{f}(x)$ will be learned in order to determine a boundary between tumor classes. A linear decision function composed by a hyperplane is one of the most used functions in classification tasks [34]. These classifiers known as *separating hyperplanes* are built to separate the samples x between classes as well as possible based on the sign of

$$f_w(x) = w^T x + b \tag{3.10}$$

where the vector w^T weights each dimension or feature of the vector x and b is a bias term. The normal vector to the surface of the hyperplane that separate the classes is defined as $w^* = \frac{\omega}{\|w\|}$. Any sample x_0 lying on the hyperplane verifies $w^T x_0 + b = 0$ and the relative distance $t(w, b)$ of any sample x to the separating hyperplane is determined as $t = \frac{1}{\|w\|}\left(w^T x + b\right) = \frac{1}{\|w\|} f(x)$. The distance t from a sample x to the separating hyperplane has a positive or negative sign depending on which side of the hyperplane lies the x sample [32].

Support Vector Machines

The main classification model used in this thesis is the Support Vector Machine (SVM) classifier introduced by Vapnik and Chervonenkis [35]. The SVM model has been widely used in different domains like computer vision, signal processing and bioinformatics. To introduce and formalize the SVM model first the Optimal Separating Hyperplanes have to be defined. Given a set S of n labeled samples (x_i, y_i) with $i = 1...n$ where $x_i \in \mathcal{X}$ and $y = \{-1, 1\}$ a partition $D(x)$ of $\mathcal{X}$ whose border is the hyperplane defined by all S so that $w^T x + b = 0$.

The partition $D(x)$ is a decision rule defined as the sign of the orthogonal projection of x on the vector normal to the hyperplane as

$$\begin{array}{rl} D(x) = & \text{sign}[w^T x + b] \\ D(x) = & \text{sign}(f(x)) \end{array} \tag{3.11}$$

If the two classes $\{-1, 1\}$ are linearly separable it is possible to learn a function $f(x)$ such that

$$\begin{array}{lll} w^T x_i + b & \geq 1 & \text{if} \quad y_i = 1 \\ w^T x_i + b & \leq -1 & \text{if} \quad y_i = -1 \end{array} \tag{3.12}$$

which can be restated as

$$y_i f(x_i) \quad = y_i(w^T x_i + b) > 1 \quad \forall\, i \tag{3.13}$$

then keeping the context of learnly separable classes they may cast an infinity number of hyperplanes that satisfies (13). Among those Vapnik and Chervonenkis suggested, as induction principle, to select the one that maximizes the distance between the two classes. That distance known as the *margin* M and equal to $\frac{2}{\|w\|}$. The problem is to maximize $\frac{2}{\|w\|}$ which can be modeled as the following optimization problem [35].

$$\begin{array}{c} \min\limits_{w,b} \|w\| \\ \text{s.t.} \quad y_i(wx_i + b) \geq 1 \quad \forall\, i = 1...N \end{array} \tag{3.14}$$

In the case where the classes overlap or are not linearly separable the former problem has no solution. To tackle that problem a set of slack variables $\xi = \{\xi_1, \xi_2, ..., \xi_n\}$ can be introduced to allow samples to violate the strict constraint introduced in equation 14. The new constraints introduce the concept of *soft margin* and can be formalized as

$$\begin{array}{c} y_i(w^T x_i + b) \geq (1 - \xi_i) \\ \forall i \ \xi_i \geq 0 \end{array} \tag{3.15}$$

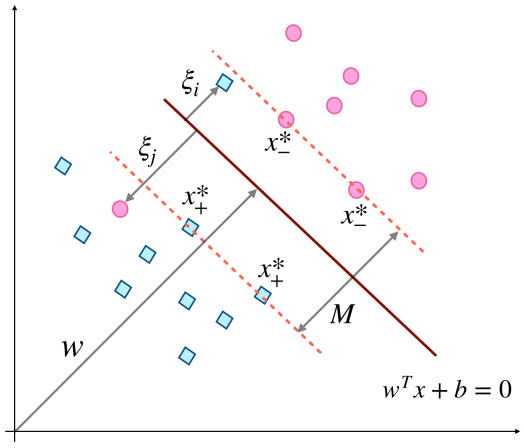

Figure 3.3: Binary classification problem with soft-margin support vector machine classifier. M is the margin, x^* support vectors of $+$ and $-$ classes and ξ are the slack variables for samples lying on the wrong side of the decision boundary.

Figure 3.3 shows an example of a binary classification problem using the soft-margin support vector machine classifier. The minimum value of ξ_i is proportional to the amount of the constraint violation between $y_i f(x_i)$ and the strict application of the constraint in equation 15. The objective is to minimize the violation importance of the $\sum \xi_i$ that represents the total error score of samples that lies in the wrong side of the margin and to maximize the margin. Only samples with a slack variable $\xi_i > 1$ are the ones in the wrong side and are named *misclassifications*. The optimal separating hyperplane in non linearly separable classes is learned by solving an optimization problem to find the direction w in the form of

$$\min \|w\|^2 \text{ s.t. } \begin{cases} y_i(w^T x_i + b) & \geq (1 - \xi_i) \\ \xi_i \geq 0 \end{cases} \tag{3.16}$$

which is a convex optimization problem. Well classified samples do not play an important role in the definition of the hyperplane. The samples x_i lying just on the margin with $y_i(w^* x_i + b) = 1$ and those on the wrong side of the margin with $\xi_i > 1$ are the ones that defines the final boundary and are called support vectors.
To find the optimal w vector a quadratic programming problem is stated as

$$\begin{gathered} \min_{w,b} \tfrac{1}{2}\|w\|^2 + C \sum_{i=1}^{N} \xi_i \\ \text{subject to } \xi_i \geq 0, y_i \left(w^T x_i + b\right) \geq 1 - \xi_i \quad \forall\, i \end{gathered} \tag{3.17}$$

where the parameter C (the cost) represents the importance given to constraint violation. The quadratic programming problem of equation 17 is solved by introducing Lagrange Multipliers α and by maximizing the dual objective function of the lagrangian [32] as

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i^T, x_j \rangle$$
$$\alpha \in [0, C] \tag{3.18}$$

Finally the solution w is defined as

$$\hat{w} = \sum_{i=1}^{N} \alpha_i y_i x_i \tag{3.19}$$

and $\alpha_i > 0$ only for a reduced subset of samples x_i named *support vectors*. The solution w depends only in the support vectors and these can lie exactly on the margin if $\xi_i = 0$ and $0 < \alpha_i < C$ or can be on the wrong side of the margin with $\xi_i > 0$ and $\alpha_i = C$. Once the w vector is obtained the decision function $D(x)$ of equation 5 is obtained. The only hyper-parameter to tune is the cost C.

Example:

To visualize an example of the SVM classifier on real data 135 samples of lung cancer from the ICGC data repository have been used to train a model. Train a model means finding the optimal separating hyperplane between overlapped classes. The classes are two lung cancer subtype: Squamus Cell and Adenocarcinoma. The number of genes used in this example are just two in order to ease the visualization. The genes are the *RPLP1* and *GAPDH* and are the ones which expression values present the highest variance in the lung cancer dataset. Figure 3.4 shows a scatter plot of the tumor samples characterized by the two genes. A SVM model has been trained to classify both tumor subtypes and a decision hyperplane with soft margin obtained. The support vectors are highlighted in red. Figure 3.4 also shows the estimated density $w^* + b$ of tumor samples for each class. It is observed that the classes are not linearly separable by using just the two observed genes causing that samples are lying significantly on the margin or on the wrong side of the separating hyperplane.

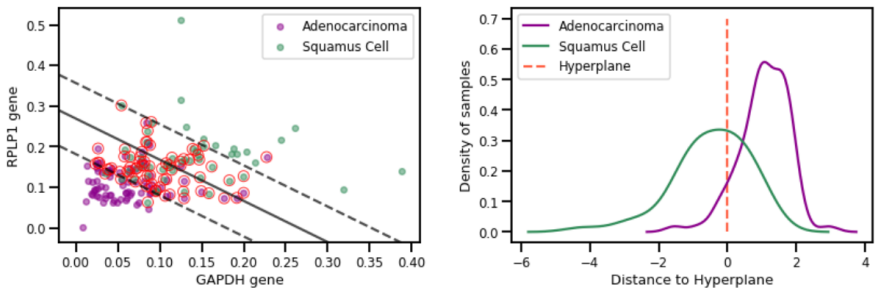

Figure 3.4: Left: Maximum margin decision function with overlapped classes. Right: Density of tumor samples of each class by considering the distance to hyperplane.

This example leads to a misslcassification rate of 20% and suggests that just two gene features and a linear model are not able to predict accurately the class of newcomers. In the real case scenario it is also considered a multiple dimension input data with $d >> 2$ and non separable classes. To tackle this scenario evaluation metrics and kernel approaches are presented in section 2.2 and 3 respectively. Evaluation metrics are required to analyze the performance of each classifier and to select the best model while Kernel methods are used to tackle non-linear problems.

3.2.2 Evaluation metrics in classification

To measure the quality of the binary decision function $D(x)$ different metrics can be computed such as classification accuracy or the area under the ROC curve.
Each binary classification problem has two classes: a Positive and a Negative class. To evaluate a classification model four outcomes are possible: True Positive (TP), True Negative (TN), False Negative (FN) and False Positive (FP). Given the Positive class the TP counts the number of samples well classified while the FN counts the positive samples classified as negative by the model. Given the Negative class the TN counts the number of samples well classified while the FP counts the miss classified ones.

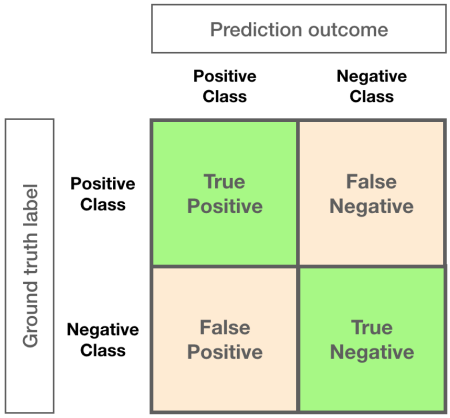

Figure 3.5: The confusion matrix. Rows: Grownd truth labels. Columns: prediction outcome.

With the four possible outcomes the Confusion Matrix is built. Confusion matrix gives an idea of how samples are distributed among all the possible outcomes of a classifier. The more samples in the diagonal of the matrix the better the classification performance. From the confusion matrix the *accuracy* score can be computed as

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TF} + \text{TN} + \text{FN}} = \frac{\text{TP} + \text{TN}}{\text{total \# of samples}} \tag{3.20}$$

and represents the proportion of samples well classified.
Another way to measure the performance of a classifier is to estimate the Area Under the

Receiving Operation Characteristic Curve (ROC) [36] or Area Under the Curve (AUC). The ROC represents the TP rate as a function of the FP rate as detailed in figure 3.6. The AUC is a score used to measure the overall performance of a binary classifier. It ranges from 0.5 to 1. An AUC = 0.5 represents the performance of a random classifier while an AUC = 1 corresponds to a perfect classifier. Unlike Accuracy or other classification metrics based on a specific decision threshold defined a priori, the AUC is independent of the classification threshold and is a measure of a global performance. This makes the AUC a robust score.

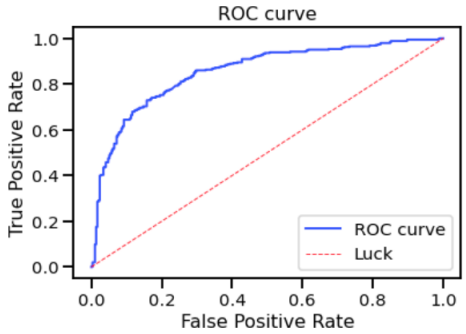

Figure 3.6: The Area Under the ROC Curve.

The presented classification performance metrics are used in the experimental section as a downstream task after feature selection to evaluate classifiers trained on different feature subsets. The performance of the classifier can be used as a quality measure of the selected feature subset.

3.3 Kernel Methods

The classification method explained previously is linear with regard to the input data . Nevertheless optimal border between the classes of genomic data may not be linear. For this reason linear models may be of limited interest in finding useful rules to classify tumor samples as detailed in Figure 3.4. To deal with non-linear classification models Kernel Methods are introduced and studied in this thesis.
Conceptually kernel methods are based on the property of certain multivariate functions to be a dot product in an unknown latent space. Kernel methods consist in a nonlinear *feature map* ϕ from the input space $\mathcal{X}$ to a high dimensional space known as *embedding or feature space embedding* $\mathcal{H}$. Then a model that learns a linear function can be trained in the feature space and back to the initial space the trained function is not linear in most cases.
Since SVM only depends on dot products it is possible to extend SVM by introducing kernels so that the model perform non linear classification. By doing so the SVM model will define a linear classifier in the feature space, but back in the original space the border is not linear anymore [18].

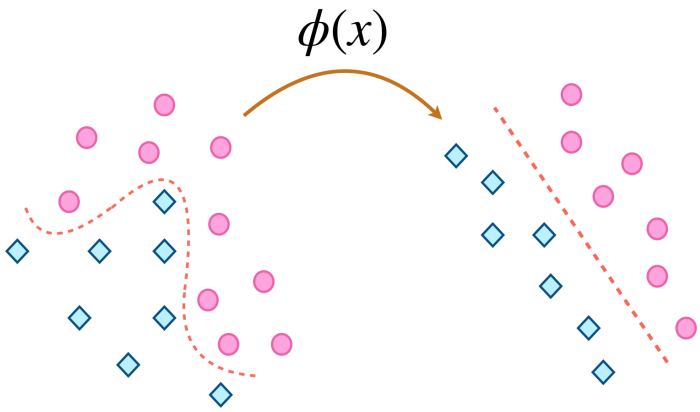

Figure 3.7: A non-linear feature map with non-linearly separable classes from the input space (left) allow the application of linear functions on the feature space (right).

To define a kernel function first inner product space and Hilbert space must be formalized. A space $\mathcal{H}$ is an inner product space between vectors $x \in \mathcal{X}$ if there exists a mapping $\phi(x)$ defined as

$$\langle \phi(x), \phi(x) \rangle \geq 0 \tag{3.21}$$

which is symmetric, bilinear and $\mathcal{H} \times \mathcal{H} \to \mathbb{R}$. A *Hilbert Space* $\mathcal{H}$ is an inner product space which is also *complete* and *separable*.
Given $\mathcal{X} \subseteq \mathbb{R}^d$ a d-dimensional space where ϕ is a mapping function from $\mathcal{X}$ to a high dimensional feature Hilbert space $\mathcal{H}$ such that

$$\begin{aligned} \mathcal{X} &\mapsto \mathcal{H} \\ x &\mapsto \phi(\mathcal{X}) \end{aligned} \tag{3.22}$$

and the dataset S of n labeled samples $S = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, where $x_i \in \mathcal{X}$ and $y_i \in \{-1, +1\}$ then S can be re-expressed as

$$S = \{(\phi(x_1), y_1), (\phi(x_2), y_2), ..., (\phi(x_n), y_n)\} \tag{3.23}$$

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a **kernel** if there is a feature map $\phi : \mathcal{X} \mapsto \mathcal{H}$ and a Hilbert Space $\mathcal{H}$ such that for any x_i, x_j there is

$$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}} \tag{3.24}$$

where Mercer theorem states that necessary and sufficient condition for k to be a kernel is that it has to be symmetric and positive semi-definite [37].

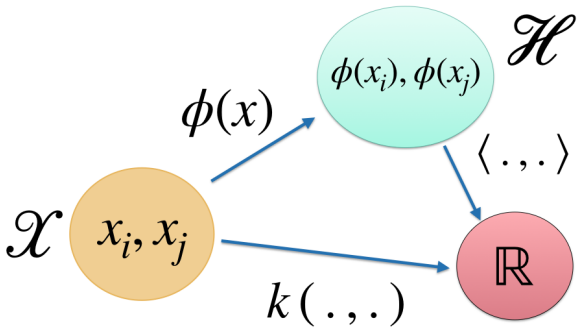

Figure 3.8: The kernel function k computes directly the pair-wise inner product in the feature space $\mathcal{H}$ without computing explicitly the feature map ϕ.

The function k computes the inner product of a pair of samples vectors in $\mathcal{H}$ the Hilbert Space. The Kernel Matrix or Gram Matrix K is defined as a $n \times n$ matrix with entries K_{ij}. Every entry of the Kernel or Gram Matrix is defined as

$$K_{ij} = \langle \phi(x_i), \phi(x_j) \rangle = k(x_i, x_j) \tag{3.25}$$

and has the following structure

K	1	2	$\cdots$	n
1	$k(\mathbf{x}_1, \mathbf{x}_1)$	$k(\mathbf{x}_1, \mathbf{x}_2)$	$\cdots$	$k(\mathbf{x}_1, \mathbf{x}_N)$
2	$\kappa(\mathbf{x}_2, \mathbf{x}_1)$	$k(\mathbf{x}_2, \mathbf{x}_2)$	$\cdots$	$k(\mathbf{x}_2, \mathbf{x}_N)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	$k(\mathbf{x}_n, \mathbf{x}_1)$	$k(\mathbf{x}_n, \mathbf{x}_2)$	$\cdots$	$k(\mathbf{x}_n, \mathbf{x}_n)$

(3.26)

The gram matrix K is symmetric then $K_{ij} = K_{ji}$ and $K = K^T$. A matrix $\mathbf{K}$ is positive semi-definite if it can be decomposed by eigen-decomposition as

$$K = \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \upsilon \upsilon^T \tag{3.27}$$

where λ_i and υ_i are the eigenvalues and the eigenvectors respectively and all the eigenvalues are nonnegative $\lambda_i \geq 0$ [38].

Example

Given a two dimensional input space $X \in \mathbb{R}^2$ with a feature map

$$\boldsymbol{\phi} : \mathbf{x} = (x_i, x_j) \longmapsto \boldsymbol{\phi}(\mathbf{x}) = \left(x_i^2, x_j^2, \sqrt{2} x_i x_j \right) = \mathbb{R}^3 \tag{3.28}$$

where the *feature map* ϕ takes two-dimensional input samples and map these to a three dimensional space by making the feature relations in the feature space linear while the input relations in the input space are quadratic. The inner products in the future map can be expressed as follows

$$\begin{aligned}\langle \phi(\mathbf{x_i}), \phi(\mathbf{x_j}) \rangle &= \left\langle \left(x_{i1}^2, x_{i2}^2, \sqrt{2} x_{i1} x_{i2} \right), \left(x_{j1}^2, x_{j2}^2, \sqrt{2} x_{j1} x_{j2} \right) \right\rangle \\ &= x_{i1}^2 x_{j1}^2 + x_{i2}^2 x_{j2}^2 + 2 x_{i1} x_{i2} x_{j1} x_{j2} \\ &= (x_{i1} x_{j1} + x_{i2} x_{j2})^2 = \langle \mathbf{x_i}, \mathbf{x_j} \rangle^2 \end{aligned} \tag{3.29}$$

This shows that it is possible to compute the inner product of sample vectors $(x_{i1} x_{j1} + x_{i2} x_{j2})^2 = \langle \mathbf{x_i}, \mathbf{x_j} \rangle^2$ via the mapping ϕ without computing explicitly their coordinates in the feature space. Then if a kernel function k meets the requirements of positive semi-definiteness and symmetry it is not necessary to compute explicitly the mapping ϕ and it is enough to express and compute only the inner product between input sample vectors in the feature space as $\langle \phi(\mathbf{x_i}), \phi(\mathbf{x_j}) \rangle$. For this reason a Kernel function can map the input features to a high dimensional Hilbert space without computing it explicitly.

The kernel function can be thought as an information bottleneck that concentrates the information required to perform a learning task [18]. A kernel acts as an interface between the input data and a learning model such as a linear classifier. Kernels can be thought as similarity functions between pair of input vectors where an output close to 0 means orthogonal samples or high dissimilar while on the other side a large output means similar samples in the Hilbert space.
In a supervised problem when the tumor labels y are available given two sample vectors belonging to the same tumor class C_1 an ideal kernel may map both samples close one to another if

$$k(x_i^{(C_1)}, x_j^{(C_1)}) \quad \simeq \quad 1 \qquad \forall x_i, x_j \in C_1 \tag{3.30}$$

In the opposite situation when a pair of sample vectors belong to different classes C_1 and C_2 respectively then an ideal kernel may map these samples in an almost orthogonal subspaces so that

$$k(x_i^{(C_1)}, x_j^{(C_2)}) \quad \simeq \quad 0 \qquad \forall x_i \in C_1 \quad , \forall x_j \in C_2 \tag{3.31}$$

Supervised problems with real labeled data such as tumor classification implies non linear and complex scenarios. In this cases it may be difficult for a kernel function to satisfy completely the two scenarios detailed above at equations 30 and 31. Nevertheless given a set of kernels, it is expected that the ones which approach the best to an ideal supervised kernel are going to perform better for classification than others.

3.3.1 Kernel Alignment

To compare different kernels the *Kernel Alignment* score is introduced. Given two valid kernels matrices G and M over a set of N samples Kernel Alignment makes it possible to determine how similar they are [39]. To compute the Kernel Alignment first the *Frobenious Inner Product* $\langle ., . \rangle_F$ between pair of kernel matrices K_1 and K_2 must be defined as

$$\langle \mathbf{K}_1, \mathbf{K}_2 \rangle_F = \sum_{i,j=1}^{N} \mathbf{K}_{1ij} \times \mathbf{K}_{2ij} = \mathrm{tr}(\mathbf{K}_1^T, \mathbf{K}_2) \tag{3.32}$$

Then the similarity between two kernel matrices $\mathbf{K}_1$ and $\mathbf{K}_2$ is expressed by the alignment A score defined as the normalized Frobenious Inner Product:

$$A\left(\mathbf{K}_1, \mathbf{K}_2\right) = \frac{\langle \mathbf{K}_1, \mathbf{K}_2 \rangle_F}{\sqrt{\langle \mathbf{K}_1, \mathbf{K}_1 \rangle_F \langle \mathbf{K}_2, \mathbf{K}_2 \rangle_F}} \tag{3.33}$$

It measures the similarity between the two kernels using the same sample set S [40][41]. The alignment can be thought as how similar both kernels maps the samples from $\mathcal{X}$ to $\mathcal{H}$. If tumor labels y are used, then a supervised ideal kernel K_{yy} can be computed as

$$K_{yy} = \begin{cases} K_{yy} = +1 & y_i = y_j \\ K_{yy} = 0 & y_i \neq y_j \end{cases} \tag{3.34}$$

and the K_{yy} matrix is known as *target kernel*. If the K_{yy} matrix shows the sample labels sorted by class then the matrix should reveal the following structure

K_{yy}	$y=-1$	$y=-1$	$\cdots$	$y=1$	$y=1$
$y_1=-1$	1	1	$\cdots$	0	0
$y_1=-1$	1	1	$\cdots$	0	0
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
$y_1=1$	0	0	$\cdots$	1	1
$y_1=1$	0	0	$\cdots$	1	1

(3.35)

From the definition of Kernel Alignment, if K_2 represents the target kernel K_{yy} the alignment of a kernel K built on x and the target K_{yy} can be expressed as

$$A\left(K, K_{yy}\right) = \frac{\langle K, K_{yy} \rangle_F}{\sqrt{\langle K, K \rangle_F \langle K_{yy}, K_{yy} \rangle_F}} \tag{3.36}$$

known as *Kernel Target Alignment* (KTA) score. The higher the KTA between a given kernel matrix K and its target K_{yy} the higher the inter-cluster separation between the two classes. The lower bound of the KTA metric is 0 since both matrices K and K_{yy} are positive semi-definite and $\langle K, K_{yy} \rangle \geq 0$. The upper bound of the KTA metric is 1 and is obtained when to $K = K_{yy}$. A value close to 1 means a high alignment between kernel matrices and it is desired in supervised problems.

3.3.2 Combination of Kernels

Real world problems sometimes are difficult and too complex to be solved with simple kernels. For this reason it is possible to combine simple kernels and obtain more complex ones while preserving the positive semi-definite and symmetry properties. Kernels can be combined by different operations. The sum of two kernels k_1 and k_2 is a valid kernel

$$k\left(\mathbf{x}_1, \mathbf{x}_2\right) = \quad k_1\left(\mathbf{x}_1, \mathbf{x}_2\right) + k_2\left(\mathbf{x}_1, \mathbf{x}_2\right) \tag{3.37}$$

The product between a positive scalar $\alpha \in \mathbb{R}^+$ and a kernel is a valid kernel

$$k\left(\mathbf{x}_1, \mathbf{x}_2\right) = \quad \alpha k_1\left(\mathbf{x}_1, \mathbf{x}_2\right) \tag{3.38}$$

The product between two kernels is also a valid kernel

$$k\left(\mathbf{x}_1, \mathbf{x}_2\right) = \quad k_1\left(\mathbf{x}_1, \mathbf{x}_2\right) k_2\left(\mathbf{x}_1, \mathbf{x}_2\right) \tag{3.39}$$

The combination of kernels may be used to improve a supervised learning task like classification where the boundary between classes is complex and difficult to find with a simple kernel [42].

3.3.3 Multiple Kernel Learning

Multiple Kernel Learning (MKL) is an approach used to combine multiple simple kernels in order to build a more complex one. MKL is defined as the linear combination of multiple kernels resulting in a final one [43] and can be expressed as

$$\boldsymbol{k_\mu}\left(\boldsymbol{x}, \boldsymbol{x'}\right) = \sum_{i=1}^{d} \mu_i k_i\left(\boldsymbol{x}, \boldsymbol{x'}\right) \ , \ \mu_i \geq 0 \tag{3.40}$$

where the vector parameter μ corresponds to the weight $\mu_i > 0$ of each kernel k_i and it is directly related to the importance of each kernel in the final solution. Despite there are different objective functions to calculate the weights of the MKL model, like the classification accuracy of a support vector machine, in this work the MKL model is built with the objective to maximize the KTA. This means that the resulting kernel from the combination of the initial ones will present a higher inter-class separability than each individual kernel on its own. The resulting KTA for the kernel matrix $\boldsymbol{K_\mu}$ is estimated by $A\left(\boldsymbol{K_\mu}, K_{yy}\right)$ and computed as detailed in equation 35.

MKL solver:

This work uses a greedy strategy [44] to compute the weights μ of the resulting $\boldsymbol{k_\mu}$ kernel by combining just two kernels at each iteration while maximizing the KTA of the resulting kernel. This strategy allows a simpler computation of the μ vector by obtaining the weights values where the derivative of the alignment becomes $= 0$ as optimal condition for each

partial derivative. Solving the MKL process with two kernels $[K_1, K_2]$ at each iteration t means that the weight vector is $\boldsymbol{\mu} = [\mu_1, \mu_2]$. Since the solution with two kernels is convex [45] the values of μ_1 and μ_2 can be obtained from

$$\begin{cases} \frac{\partial \boldsymbol{A}(K_{yy}, \mu_1 K_1 + \mu_2 K_2)}{\partial \mu_1} = 0 \\ \\ \frac{\partial \boldsymbol{A}(K_{yy}, \mu_1 K_1 + \mu_2 K_2)}{\partial \mu_2} = 0 \end{cases} \tag{3.41}$$

subject to $\mu_1, \mu_2 \geq 0$. From a set D of d kernels and the dataset S, the greedy approach starts at iteration $t = 0$ by choosing the kernel matrix K_i from D with highest KTA_i as $A(K_i, K_{yy})$. Then assign K_i to a empty list P containing the kernels which are part of the solution and makes it the current solution $K_\mu^{(t=0)} = K_i$. Then for the following iterations $t = 1...p$ the method combines iterativelly the current $K_\mu^{(t)}$ and a new kernel K_j from D with $i \neq j$ to obtain a new solution $K_\mu^{(t+1)} = \mu_\alpha K_\mu^{(t)} + \mu_\beta K_j$ that maximizes the overall KTA as $A(K_{yy}, K_\mu^{(t+1)})$. This process prioritizes in adding first to the solution P the kernels that increase the most the overall KTA, thus the kernel selected first will have the highest weight μ_i. The following kernels to be selected will have a lower weight μ_i since their contribution to maximize the overall KTA as ΔKTA is going to be lower than the previous added kernel. The KTA increment at each iteration will be $\Delta KTA^{(t)} > \Delta KTA^{(t+1)} ... > \Delta KTA^{(p)}$. This means that when a number p of kernels is requested to be in the solution weights tend to decrease with p as $\mu_0 > \mu_1 > .... > \mu_p$. Despite the MKL optimization problem presents a global solution the used greedy strategy with a fixed p number of kernels to be selected in the linear combination is is suboptimal but efficient when the number of considered kernels is large [44].

3.3.4 Types of Kernels

There are multiple functions that meet the kernel properties detailed previously. These functions compute the inner product of two samples mapped into the Hilbert space without the explicit computation of it.
The linear kernel is defined as the dot product between sample vectors as

$$k_{\text{lin}}(\mathbf{x_i}, \mathbf{x_j}) = \langle \mathbf{x_i}, \mathbf{x_j} \rangle \tag{3.42}$$

The Polynomial kernel is

$$k_{\text{pol}}(\mathbf{x_i}, \mathbf{x_j}) = (\langle \mathbf{x_i}, \mathbf{x_j} \rangle + b)^r \tag{3.43}$$

where b and r are kernel parameters. The Radial Basis Function (RBF) Kernel is defined as

$$k_{\text{rbf}}(\mathbf{x_i}, \mathbf{x_j}) = \exp\left(-\gamma \left\| \mathbf{x}_i, \mathbf{x}_j \right\|^2\right) \tag{3.44}$$

where γ is a scale parameter. The RBF kernel is a generalization of the Gaussian Kernel

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{z}\|^2}{2\sigma^2}\right) \tag{3.45}$$

where $\gamma = \frac{1}{2\sigma^2}$. The higher the γ means the lower the σ parameter then the more sensitive to details will be the kernel and high frequency functions will be captured. On the opposite for low γ means a high σ thus the kernel will be more simple and less sensitive to details. The RBF Kernel is known to be a good kernel candidate in any application and is the function used in the experiments of this thesis [46].

Example

Following with the Lung Cancer example presented previously with two classes based in tumor subtypes and characterized by two genes RPLP1 and GAPDH the kernel Gram matrices are built. The Kernel matrices learned from this data are detailed in figure 3.9 and are built from RBF kernels each one with different values of γ. Additionally the target kernel matrix K_{yy} built with the labels is detailed as explained in equation 34.

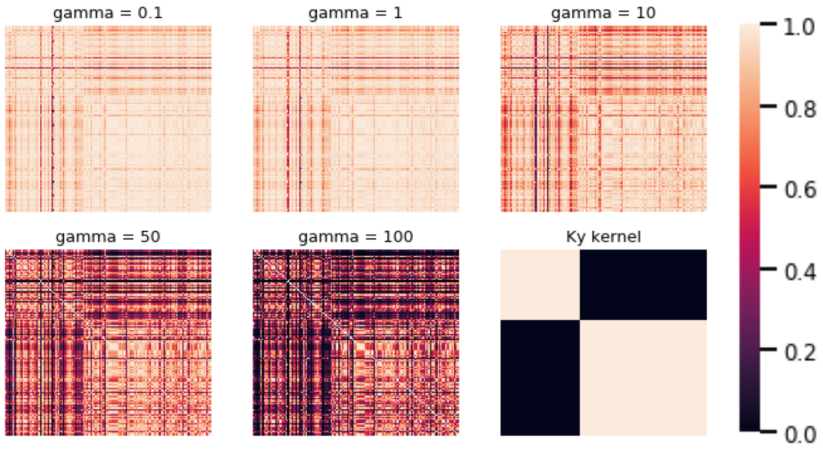

Figure 3.9: Gram matrices for RBF kernels with different values of γ and the target kernel built from labels.

The K_{yy} target matrix shows that there are two classes ideally distributed each one in one cluster separately. In supervised learning problems K_{yy} is the target matrix and it is expected that every kernel learned from training data aligns as much as possible to K_{yy} since it implies the greatest separability between classes. The matrices showed in Figure 3.9 shows with intensity the similarity between each pair of samples from the training data. Values of the γ parameter $10 < \gamma < 100$ reveals a structure in the resulting gram matrices closer to K_{yy} than for smaller values. So the two classes are more visible for this range of values.

It is possible to evaluate numerically the quality of each kernel from a supervised learning perspective by computing the KTA of each kernel with respect to K_{yy}. Table 1 shows the resulting KTA score for the RBF kernel across different values of $\gamma = [0.1, 1, 10, 50, 100]$. The highest KTA is obtained with $\gamma = 50$.

Gamma	$\gamma = 0.1$	$\gamma = 1$	$\gamma = 10$	$\gamma = 50$	$\gamma = 100$
KTA	0.725	0.729	0.746	**0.749**	0.722

Table 3.1: Kernel target alignment for different values of γ parameter.

If the grid of values of γ is refined then it is possible to visualize a peak in the KTA value around $\gamma = 30$. Figure 3.10 shows the KTA as a function of γ.

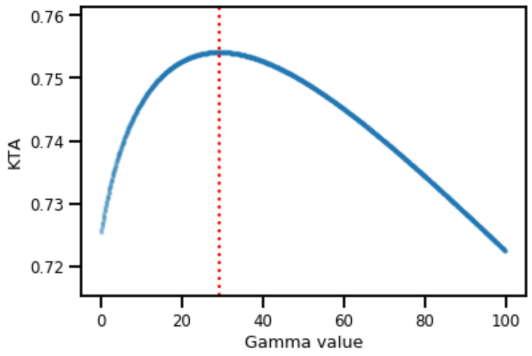

Figure 3.10: Kernel Target Alignment of a RBF Kernel as a function of the γ parameter.

This evidences that some values of γ are more convenient than others in supervised problems. The value of γ that maximizes the KTA can change depending on the structure of training data. A higher KTA means a higher inter-class separability in Feature Space $\mathcal{H}$ (Hilbert Space). For this reason it is expected that high KTA values improve classification compared to a low KTA.

3.3.5 Kernel machines in classification

The support vector classifier presented previlusly finds a linear boundary in the input feature space by defining a hyperplane with optimal separability between classes. By using kernels a new feature space can be obtained with a non-linear transformation ϕ and thus makes it possible to solve non-linear problems where classes are highly overlapped and are not linearly separable using a linear technique in the Feature Space $\mathcal{H}$.
The core idea is to apply a transformation/mapping to the input feature vector X and then use linear models in the new space. Then the dot product between sample vectors $\langle x_i, x_j \rangle$ in the dual Lagrange optimization problem detailed in equation 14 can be expressed as $\langle \phi(x_i), \phi(x_j) \rangle$ resulting in

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \left\langle \phi\left(x_i\right), \phi\left(x_j\right) \right\rangle \tag{3.46}$$

Then the decision function can be re-written as

$$f(x) = \langle \phi(x)^T w \rangle + \beta_0 = \sum_{i=1}^{M} \alpha_i y_i \langle \phi(x), \phi(x_i) \rangle + b \tag{3.47}$$

where $\phi(x)$ is used only for inner products. For this reason it is not necessary to determine the transformation $\phi(x)$ but it is required to know the positive and semi-definite kernel function $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ responsible for computing the inner products in the transformed space. The kernel function can be used in the decision function as

$$\hat{f}(x) = \sum_{i=1}^{N} \hat{\alpha}_i y_i k(x, x_i) + \hat{b} \tag{3.48}$$

By using a non-linear kernel the decision function detailed in equation 48 is then a non-linear boundary in the input space. The benefit of this characteristic is at least better classification accuracy in the train set and at most a better classification rule.

Example:

By training a SVM classifier with a RBF (Gaussian) kernel on the Lung dataset used previously it can be seen in Figure 6 how the decision boundary is nonlinear in the input space. As consequence the number of well classified tumor samples is higher.

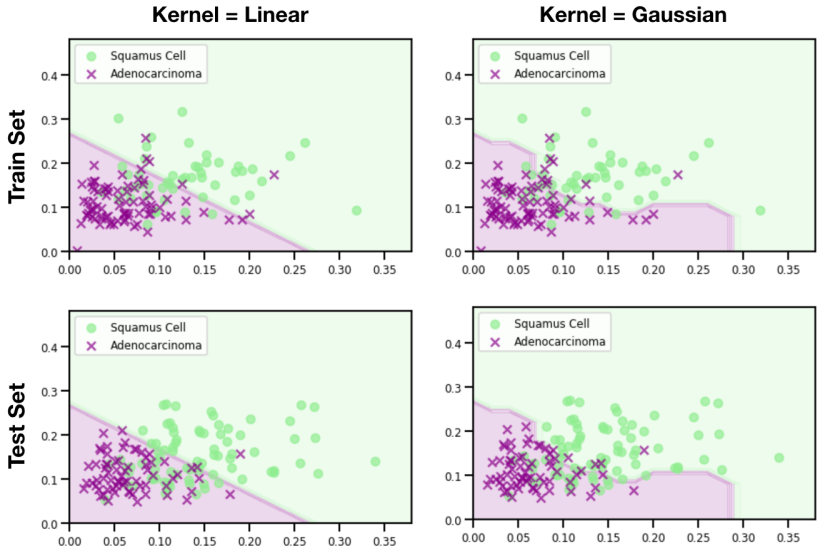

Figure 3.11: SVM classifier using linear and RBF kernels on train and test sets of Lung cancer data.

Figure 3.11 results are detailed in Table 3.2 and shows how by a Gaussian Kernel the

classification boundary can better fit the training data than the linear case and in consequence generalize better to unseen samples in test set.

Accuracy	Linear kernel	RBF kernel
Train	0.79	0.82
Test	0.73	0.76

Table 3.2: Accuracy of SVM classifiers using linear and non linear kernels.

3.4 Supervised Feature Selection

As seen in chapter 02 the gene expression data is high dimensional since it is characterized by tens of thousands of gene expression features. Moreover, the number of samples in the tumor datasets are not enough to decribe and fill the high dimensional input space. This carries out multiple problems. The first one is the cost that implies from a biomedicine perspective to measure tens of thousands of features. From the statistical learning perspective the problem is about the curse of dimensionality where the dataset has a sample to feature ratio $r << 1$, a complex scenario for learning algorithms. Finally, not all the initial input features are assumed to be useful to predict the tumor labels y, an important number of features may be noisy, irrelevant and redundant and a small fraction is assumed to contain signal information. This assumption suggests that supervised learning models can perform better if they are trained using only a small fraction of the input features, commonly named as informative or relevant features.

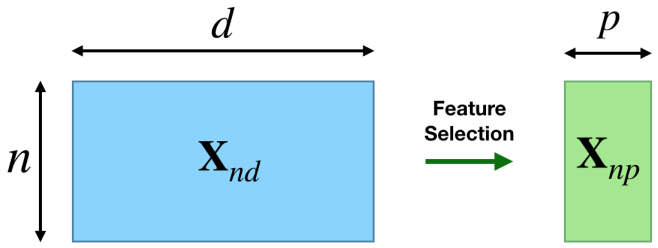

Figure 3.12: Size of the data matrix $\mathbf{X}$ before (left) and after (right) feature selection.d is the number of input features and p is the number of selected features.

Feature Selection methods have the goal to select a reduced subset of p informative features from the d input ones where $p << d$. It is expected that a supervised learning task will perform better in classification if it is trained with p features than if it is trained with the original d features since the sample density increases which gives a more significant description

of the data distribution. As a consequence it reduces the uncertainty when searching for the decision function $f(x)$.
There are three types of features: redundant, irrelevant and relevant. Redundant features are relevant but they are highly correlated between each other and explain the same information about the distribution of the data. Redundant features live in a smaller subspace so only a subset of these features is necessary to get the corresponding information. Irrelevant features do not bring enough or any information about the distribution of the data and can be considered as noise. Finally, relevant features explain the necessary information about the data distribution. Sometimes single features considered alone do not keep enough information and may be identified as irrelevant, nevertheless these can be considered as relevant when combined with other features. An initial input feature set D may contain the three types of features. The feature selection process aims to retain only the relevant ones.
Figure 3.13 shows a toy example with synthetic data of 500 samples characterized by just two features $[x_1, x_2]$. The left image shows how the features x_1 and x_2 are highly correlated which means that just one feature is enough to understand the data distribution. The center image shows how the data is mainly distributed across x_1 and almost without any variability across x_2 suggesting that the latter is irrelevant. Finally the right figure shows the data distributed between two clusters where one of these has high variability across x_1 and the other across x_2 suggesting that both features are necessary and relevant to understand the data distribution.

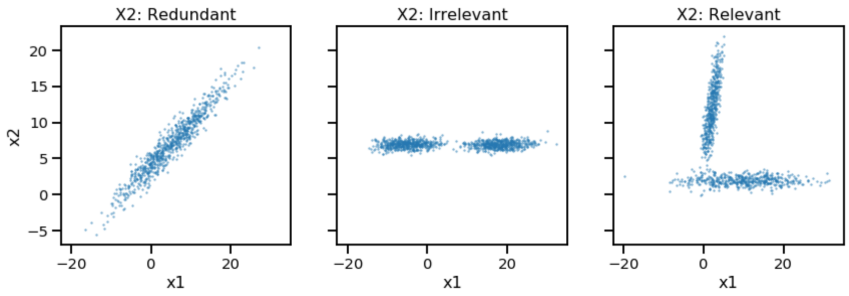

Figure 3.13: Left: x_2 feature is redundant with x_1. Center: x_2 is irrelevant and only x_1 explain the data distribution. Right: both x_1 and x_2 are necessary to explain the data distribution then both are relevant features.

By keeping the relevant features it is expected to improve the classification performance of a supervised learning algorithm like support vector classification.

3.4.1 Supervised Feature Selection methods

To perform feature selection and reduce the dimensionality of data different strategies have been developed since an exaustive search is imposible in most of the cases due to the fact there are $2^d - 1$ possible selections. The feature selection strategies can be categorized as filter, wrapper and embedded methods [26].
The wrapper methods include the iterative and greedy algorithms built to select features based

on their contribution in a classification model which guides the selection at each iteration. The classifier is trained repeatedly with different feature-subsets with the objective to find the subset that improves the most the classification performance with the chosen learning approach. The wrapper method used in this thesis is the Recursive Feature Elimination (RFE) [47].
The embedded methods are similar to the wrapper ones since they also use a learning model to guide the selection with the difference that are not greedy and the learning model is intrinsically within the feature selection model. The embedded methods used in this thesis are Hilbert Schmidt Independence Criterion Kernelized Lasso (HSIC-Lasso) [48] and Minimal Redundancy Maximal Relevance (MRMR) [49].

3.4.2 Recursive Feature Elimination Method

The SVM-RFE is a greedy feature selection method [50][51] that generates a ranking list of features at each iteration and discard the least important feature in the ranking until a stop criterion is met which selects a subset of features. The ranking is built by a feature weight vector w obtained from the parameters of the decision function of a linear SVM classifier. At each iteration of the stepwise algorithm a SVM classifier is trained and a separating hyperplane is obtained. Since the decision function is expressed as $D(x) = \text{sign}(x^T\mathbf{w} + b)$ the weight parameter $\mathbf{w}$ associated features of the SVM hyperplane function is expressed as $\mathbf{w} = [w_1, w_2, ..., w_d]$. Then the feature ranking for each gene i is defined as $c_i = w_i^2$ and the feature with the smallest c_i is removed at each iteration [52].

3.4.3 Minimum Redundancy Maximum Relevance Method

The Minimum Redundancy Maximum Relevance feature selection method (mRMR) attempts to discard redundant features while keeping the features with highest correlation to the target labels y. The objective is to select a feature subset that best characterizes the statistical property of a target label [53]. The selection process has the constraint that selected features must be mutually as dissimilar to each other as possible while being as similar to the target label as possible.
The mRMR performs a feature ranking with the criteria of maximizing the relevance R score between selected features and target label while minimizing simultaneously the redundancy U score between the selected features. To get both the relevance R and redundancy U first the Mutual Information score between two random variables x and y is computed from their corresponding probabilistic density functions $p(x)$, $p(y)$ and $p(x, y)$ as

$$I(x; y) = \iint p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dxdy \tag{3.49}$$

Since $I(x, y)$ is difficult to estimate Peng et Al. propose to replace it by Relevance score R for each feature. The feature relevance score R is calculated as an optimization problem

according to the selected subset of features from the feature set D that maximizes the following mutual information

$$\max_{x_{(.,i)}} \mathrm{R} = \frac{1}{|D|} \sum_{x_{(.,i)} \in D} I\left(x_{(.,i)}; y\right) \tag{3.50}$$

with $y \in \{-1, 1\}$ the labels and $x_{(.,i)}$ the selected features. The mutual information in the Relevance score is used to select features that has the largest dependency with the target variable y. The redundancy score U is obtained from an optimization problem according to the selected subset of features that minimizes the mutual information or the dependency between the selected variables $x_{(.,i)}, x_{(.,j)}$ as [54]

$$\min_{x_{(.,i)}} \mathrm{U} = \frac{1}{|D|^2} \sum_{x_{(.,i)}, x_{(.,j)} \in D} I\left(x_{(.,i)}, x_{(.,j)}\right) \tag{3.51}$$

Then by combining both scores the minimal redundancy maximum relevance mRMR score $\Phi(R, U)$ is finally defined by the following min-max optimization problem

$$\max_{x_{(.,i)}} \Phi(R, U) = R - U \tag{3.52}$$

3.4.4 HSIC-Lasso Method

The Hilbert-Schmidt Independence Criterion Lasso (HSIC-Lasso) method [48] is a feature-wise kernelized Lasso that captures non-linear dependency between input features and target labels. The non-redundant features with high non-linear dependency to the output labels is determined by the independence measure based on Kernel Methods *Hilbert Schmidt Independence Criterion* (HSIC) [55].

$$\min_{\alpha_1, \ldots, \alpha_d} \frac{1}{2} \left\| \mathbf{K}_{yy} - \sum_{i=1}^{d} \alpha_i \mathbf{K}^{(i)} \right\|_F^2 + \lambda \sum_{i=1}^{d} |\alpha_i| \quad \text{s.t. } \alpha_1, \ldots, \alpha_d \geq 0 \tag{3.53}$$

The idea is to build a regression of the $\mathbf{K}_{yy}$ matrix by a linear combination of feature-wise input kernels $\mathbf{K}^{(i)}$. Then Lasso is used to determine the regression and enables to select features that contribute to predict $K_y y$. The HSIC-Lasso measurement is related to the Kernel Target Alignment metric defined previously [40]. The $\alpha_i > 0$ parameters determines the selected features.

3.4.5 Evaluation of Feature Selection Methods

This chapter thesis is focused on binary classification of tumor samples by using reduced feature subsets. In order to compare the selected subsets of features obtained with different approaches two scores are used.
The first one is known as a *downstream task* since it performs a supervised learning task such

as classification after the selection and the quality of the selected features is evaluated based on the performance of the classifier. The classification task gives an idea about the predictive power of the selected features. The AUC is used to evaluate the classification performance of a binary classifier trained on the selected features. In this thesis for all the experiments a Support Vector Classifier is trained using the selected features of each method.
To evaluate the quality of the feature selection process by itself the Redundancy Rate (RED) [56] [48] metric is used. The RED score measures the mean value of absolute correlation between features is computed as

$$\text{RED} = \frac{1}{p(p-1)} \sum_{f_i, f_j \in \mathbf{P}} |\rho_{ij}| \tag{3.54}$$

with $i \neq j$ where ρ_{ij} is the correlation score between the *ith* and the *jth* selected variables. The RED metric takes values between 0 and 1. A low value of RED indicates that the majority of the selected features within the subset P have low linear correlation between each other and thus a low redundancy is expected within the selected feature set which can be interpreted as a high quality of selection. On the other side, values of RED close to 1 correponds to a subset of features with high redundancy which is a non desired output.
Additionally, if a synthetic dataset can be generated with a known fraction of informative and noisy features then it is possible to compute the accuracy of each selection method to select the informative features. The feature selection accuracy is computed as

$$FS_{\text{acc}} = \frac{SF}{IF} \tag{3.55}$$

where SF is the number of informative selected features and IF the total of informative features. The higher the accuracy the better the model performance.

3.5 Proposed method: Kernel Latent Regularization Feature Selection

The previous sections stated the importance to reduce the dimensionality of gene expression data to perform tumor classification. Some of the benefits are the reduction of complexity, avoiding the curse of dimensionality and to gain insight about which genes are important in a classification task to guide biomarker discovery.
Most of the time, the important features are selected using a supervised objective function as detailed previously [57]. In some cases, such as lack of data, the supervised objective may be too strict and difficult to fulfill in order to obtain a model that could generalize on new unseen data [58]. In this limited scenario the proposed idea is based in considering the structure of the data as another source of learning besides the sample labels y and use it as an improvement strategy for feature selection to help focusing on meaningful feature set. Then a major question arises: is it possible to improve the feature selection process by

constraining the selected feature subspace not only to infer the targeted labels but also to retain more information about the structure of the data in the initial feature space so that overfitting in the feature selection process could be avoided?
Our work proposes a feature selection method based on Multiple Kernel Learning (MKL) [43]. Additionally the proposed method combines MKL and a nonlinear latent feature extraction model to improve the feature selection process by a combination of supervised and unsupervised approaches respectively. This combination of approaches aims to improve the generalization capacity in classification of the selected features. This strategy aims to maximize the separability between tumor classes while considering simultaneously the latent structure of the training data. The proposed selection method performs what we name a *latent regularization* using simultaneously the labels of the data and unsupervised latent variables. To extract the latent variables of the training data an unsupervised dimensionality reduction model is used, more precisely the kernel-PCA (kPCA) [59]. The key idea is to search for features that consider the tumor labels and preserve the data structure simultaneously. The obtained space should be more robust to noise and lead to better generalization while dealing with classification tasks.
The proposed method is designed to deal with tumor classification problems where dimensionality $d > 18.000$ and the sample size is lower than 200 tumor profiles. Tumor profiles are classified by stage or prognosis. In this scenario most of feature selection algorithms may under perform due to the lack of tumor samples and the high dimensional feature set. The proposed latent regularisation works as a label relaxation process that is shown to improve the tumor classification performance on new unseen test samples. The proposed method is named *Kernel Latent Regularization Feature Selection* (KLR-FS).

3.5.1 Feature Selection with MKL

To explain the KLR-FS pipeline this section formalizes first how to perform feature selection with MKL.
Given a dataset of n samples characterized by d features, this work proposes to use d feature-wise kernel Gram matrices K_i. Feature-wise kernel γ_i parameters are selected in order to maximize the alignment of each kernel with a target Gram matrix K_{yy}. Then using the proposed MKL method a subset of feature-wise kernels P represented by their p Gram matrices K_i is iteratively selected and combined to increase the overall estimated KTA between the gram matrix $\boldsymbol{K}_\mu$ and K_{yy}. Only the feature-wise kernel matrices that increases the KTA are included in the final solution $\boldsymbol{K}_\mu$. This approach leads to a sparse solution where the number of selected features p associated to the selected feature-wise kernel matrices is $p << d$. The desired output of the greedy MKL approach is a reduced set of p features and a kernel function $\boldsymbol{k}_\mu$ that improve the inter-cluster distance between samples of different tumor classes and in consequence improve the support vector classification task after feature selection. The positive values of the resulting weighting vector $\mu_i > 0$ indicates the feature importance. Figure 3.14 details how features are selected via MKL by improving the alignment with a fully supervised target kernel K_{yy}. Once $\boldsymbol{K}_\mu$ is built it is used as a

custom kernel function in a support vector classifier for binary classification [37] .

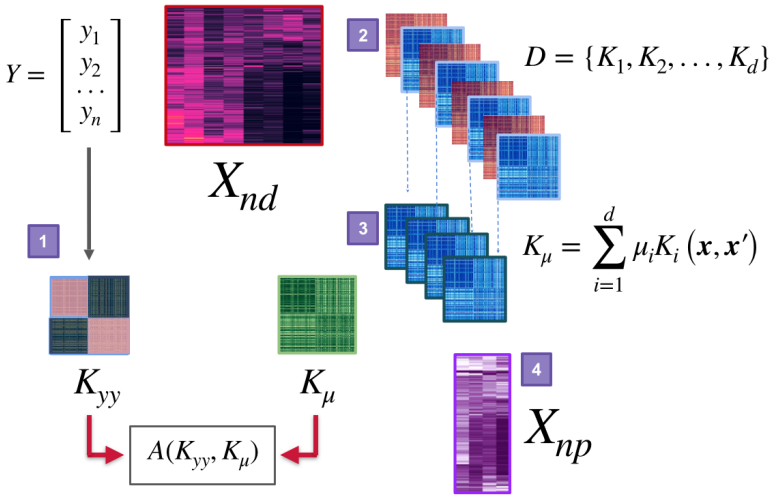

Figure 3.14: Feature selection pipeline with MKL. 1) A supervised K_{yy} gram matrix is built from tumor labels. 2) From the data matrix $\mathbf{X}_{nd}$ a set D of d feature-wise kernels is built. 3) By MKL a subset of kernels are selected by improving the alignment of the resulting kernel $A(K_{yy}, K_\mu)$. 4) The subset of p genes corresponding to the feature-wise kernels where $\mu > 0$ are selected.

3.5.2 Latent regularization with nonlinear feature extraction

In the previous sub-section a feature selection method based on multiple kernel learning and kernel target alignment is described. This method selects features using a supervised kernel matrix K_{yy} where the separation between classes is ideal as target. But this objective is very specific and since the number of samples is small it can lead to significant different solutions when changing samples in the training set. Thus it may be interesting to be less specific and more conservative about the structure of the data. For this reason a relaxation of this supervised constraint is proposed in this work by mixing the target kernel k_{yy} with a k_z kernel built on the latent space $\mathcal{Z}$ learned by a dimensionality reduction method based on a non-linear transformation $\phi_z(x)$. Note that the k_{yy} is built from the tumor labels and the k_z from the extracted latent space $\mathcal{Z}$, thus the first one is built by a supervised approach and the latter by an unsupervised one that tends to keep all information related to the general structure of the data. The mixture of both kernels forms a new target kernel k_δ that has supervised and unsupervised information from the training samples. For the unsupervised and nonlinear dimensionality reduction transformation $\phi_z(x)$ any algorithm can be used such as Kernel-PCA, Autoencoders or t-SNE. The key aspect is to learn an unsupervised latent space $\mathcal{Z}$ from the training samples in $\mathcal{X}$ to capture the structure of the data and to mix it with the supervised labels y as a mean to relax the supervised constraint.
Most supervised feature selection models target to build an inference function f that captures

the relation between features and labels expressed as

$$y = f(x)$$

where a subset p of selected features minimizes a loss function between the model output $f(x)$ and the true labels y. This work proposes a feature selection model that learns not only from the true labels y but also from the latent structure of the training data as

$$(z, y) = f(x)$$

where $z = \phi(x)_z$ is the latent space obtained from a nonlinear and unsupervised mapping function. The mixture between the labels y and z creates a hybrid target composed by a supervised and unsupervised approach.
In this work to extract a latent space z the function $\phi(x)_z$ is learned by a Kernel-PCA method [59]. This method is based on a nonlinear mapping $\phi_z(x)$ from $\mathcal{X}$ to $\mathcal{H}$ where the standard Principal Component Analysis (PCA) algorithm is performed.
The PCA is a linear dimensionality reduction method and is defined as the orthogonal projection of the training samples into a low dimensional subspace such that the variance of the projected samples is maximized [60]. From a set $S = \{x_1, ..., x_n\}$ of n samples each one characterized by d dimensions, the goal of PCA is to map x_i into a low dimensional space of dimension p such that $p < d$. The PCA projects the data into the p-dimensional subspace that maximizes the variance. The PCA transformation is obtained by applying the spectral decomposition to the covariance matrix C of the training data S_{tr} such that $\mathrm{U}\Lambda = C\mathrm{U}$. The U matrix contains stacked vectors $u_1, ..., u_p$ where u_i is the *ith* eigenvector corresponding to the *ith* largest eigenvalue λ_i in the Λ matrix. Then by selecting the first p eigen-vectors a new $n \times p$ matrix U' is obtained and works as the desired subspace. Then by computing $XU' = Z$ the original data is projected by U' into Z of dimension p.
To make PCA nonlinear the Kernel approach introduced in the previous section is used to define the Kernel-PCA in order to perform an implicit PCA in the Hilbert space $\mathcal{H}$ induced by the kernel [61]. In this case the spectral decomposition is applied to a Kernel Matrix $\mathbf{K}$ such that $n\Lambda U = \mathbf{K}U$ where U and Λ are the matrices containing the eigenvectors and eigenvalues respectively and n the number of samples. Then the resulting coordinates $(z_1, ..., z_p)$ known as kernel principal components are calculated by

$$z_j = \sum_{i=1}^{n} u_{ij} k(x_i, x) \quad , j = 1...p \tag{3.56}$$

and this is equivalent to perform a nonlinear PCA in the original input space. The number of latent variables p is an external hyperparameter.
Once the kernel-PCA is computed with the training samples $\boldsymbol{x}_{tr}$ these are mapped to the latent space $\mathcal{Z}$ and with the chosen kernel k_z a Gram Matrix $\mathbf{K}_z$ is built. The Gram Matrix $\mathbf{K}_z$ captures the unsupervised structure of the training samples in the latent space. Then by a linear combination of the supervised K_{yy} and unsupervised K_z gram matrices a new hybrid target matrix $\mathbf{K}_\delta$ is be created as

$$K_\delta = (1 - \delta)K_{yy} + \delta K_z \tag{3.57}$$

This new Gram Matrix K_δ contains a linear combination of both the supervised labels and unsupervised latent variables ruled by a new parameter named as *Mixture Coefficient* $\delta \in [0, 1]$.

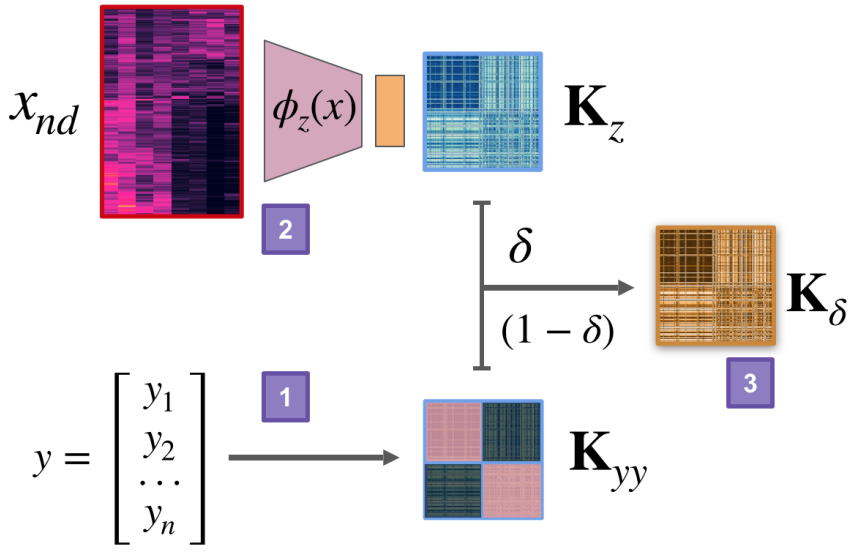

Figure 3.15: The mixture kernel matrix $\mathbf{K}_\delta$ pipeline. 1) A supervised kernel matrix $\mathbf{K}_{yy}$ is built from the tumor labels. 2) A $\mathbf{K}_z$ kernel matrix is build on the latent space $\mathcal{Z}$ learned from the kPCA. 3) A mixture kernel $\mathbf{K}_\delta$ matrix is obtained by a weighted sum of the previous kernels.

Figure 3.15 shows how the k_δ kernel and its corresponding Gram matrix $\mathbf{K}_\delta$ are built. From equation 3.57 three scenarios are possible. When $\delta = 0$ then $K_\delta = K_{yy}$ and corresponds to the supervised kernel described in the previous section and the feature selection process is completely supervised. On the other hand when $\delta = 1$ then $K_\delta = K_z$ and it represents the unsupervised structure provided by the kernel-PCA of the training samples. When $\delta = 1$ the target k_δ does not contain any supervised information and the MKL process is unsupervised. Finally, every value of $0 < \delta < 1$ corresponds to a kernel k_δ that has a mixture of supervised and unsupervised components of the problem. Then by this approach a latent regularization of the supervised feature selection problem is possible. The hypothesis stated in this work is based on the assumption that the features selected by MKL in a mixture of supervised labels and unsupervised latent variables results in a higher generalization ability and thus in higher classification performance.

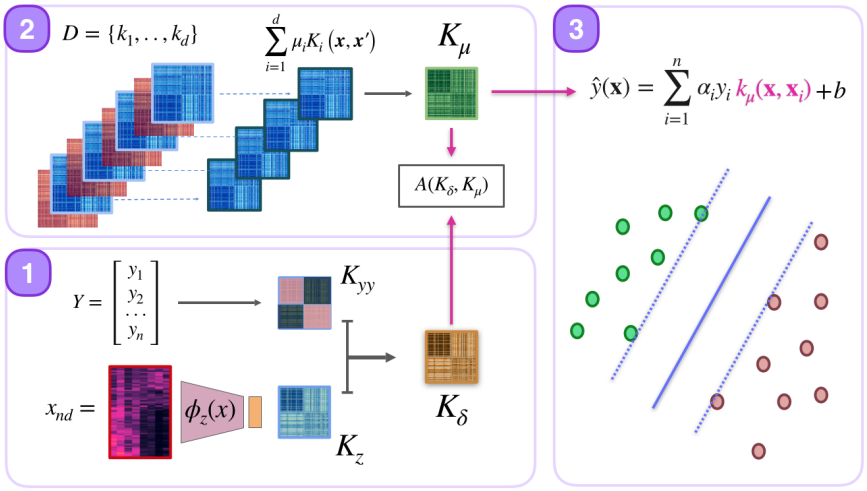

Figure 3.16: The KLR-FS pipeline. 1) A K_{yy} matrix is built using the tumor labels. A kernel-PCA model is trained using the training data and a kernel K_z is built on the latent space Z learned from the kernel-PCA. The K_δ kernel matrix is obtained from the mixture between K_{yy} and K_z matrices. 2) From the training data a set D of d feature-wise kernels is built. By MKL a subset of p feature-wise kernels are selected and a K_μ kernel matrix is obtained by improving the alingment with K_δ kernel. The μ vector indicates the selected features. 3) The k_μ kernel function is used in Support Vector Classification.

Finally, once the MKL step is done a k_μ kernel and a subset P of p selected features are obtained as an output result. With the p selected features a new subspace $\mathcal{V} \in \mathbb{R}^p$ is obtained from $\mathcal{X}$. Then using the training samples $v \in \mathcal{V}$ the k_μ kernel is used as the kernel of a support vector classifier to discriminate tumor profiles [62].

$$\hat{y}(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i \left\langle \phi_\mu(v), \phi_\mu(v_i) \right\rangle_{\mathcal{H}} + b = \sum_{i=1}^{n} \alpha_i y_i k_\mu(\mathbf{v}, \mathbf{v}_i) + b \tag{3.58}$$

Figure 3.16 details the proposed KLR-FS pipeline for supervised feature selection and Algorithm 1 presents the pseudo-code of the KLR-FS algorithm.

Algorithm 1 Kernel Latent Regularization Feature Selection algorithm

Input: X, y, p, δ
Output: K_μ, μ

Do
  Learn latent space $\phi_z(x) = Z$
  Build unsupervised latent kernel $K_z = k(z_i, z_i')$;
  Build supervised kernel $K_{yy} = \langle y, y' \rangle$;
  Build target kernel $K_\delta = \delta K_{yy} + (1 - \delta)K_z$;

  for *i from 1 to d* **do**
    Build feature-wise kernels $K_i = k(x_i, x_i)$;
    $K_i \rightarrow D$
Do
  From D select K_i where $A(K_\delta, K_i) \rightarrow$ **max**;
  $K_\mu = K_i$;
  $i \rightarrow P$;

  for *j from 1 to p* **do**
    from D select K_i with $i \notin P$ such as $A(K_\delta, \mu_1 K_\mu + \mu_2 K_i) \rightarrow$ **max** ;
    $K_\mu = \mu_1 K_\mu + \mu_2 K_i$;
    Compute μ_i
    $i \rightarrow P$
Do
  Get selected features where $\mu_i > 0$;
  Support vector classification with K_μ ;

To understand the impact of the Mixture Coefficient a toy example is presented to explore visually how the K_δ matrix is composed for different values of δ. Figure 3.17 shows a toy example using the XOR dataset with two classes and each class composed by two subclusters.

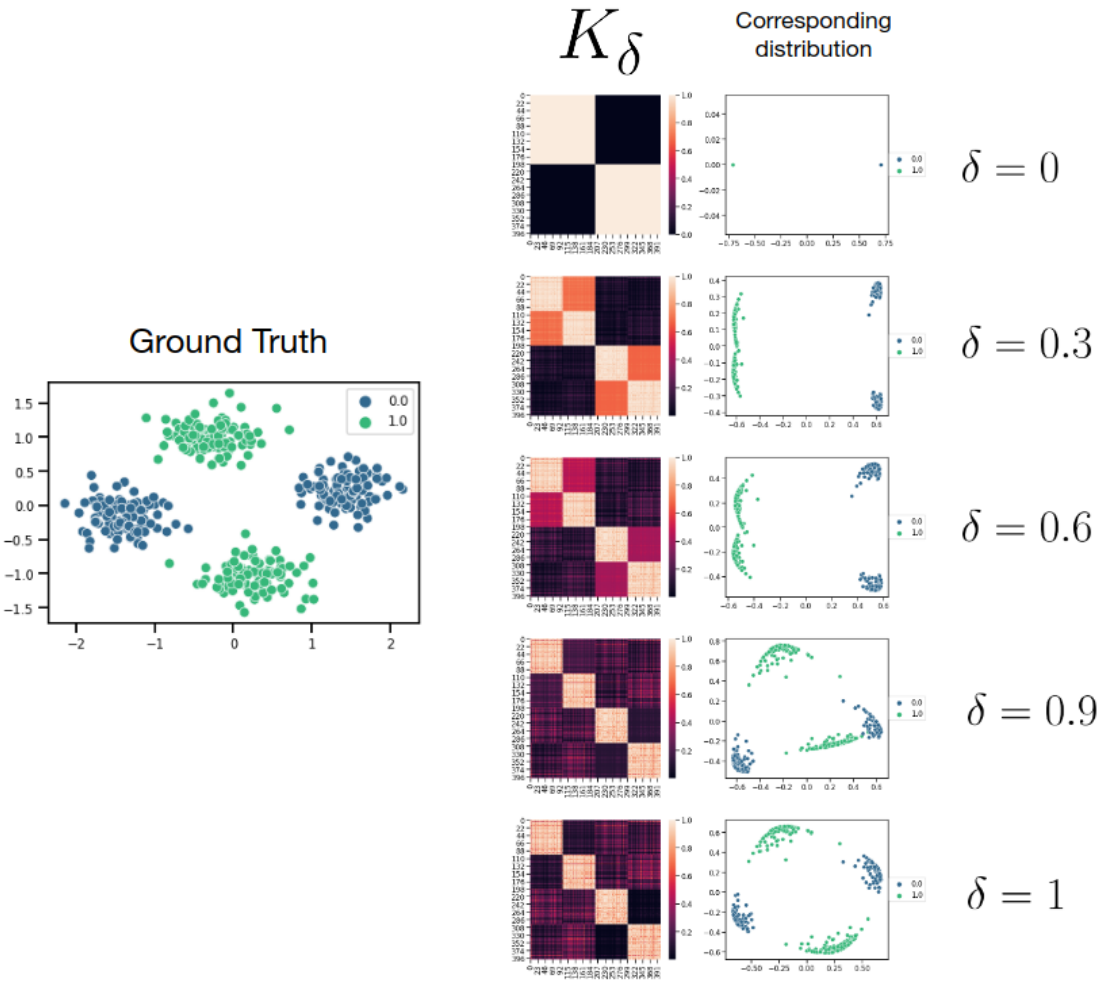

Figure 3.17: The target kernel K_δ for different values of the latent regularization.

When $\delta = 0$ then the $K_\delta = K_{yy}$ and the corresponding target distribution of the resulting kernel matrix shows just two dots for each class. This target representation is not representative of the ground truth and only contains discriminative information about the class labels but a lack of any information about the data structure. When $\delta > 0$ the latent structure of the data is mixed with the supervised one in the K_δ matrix. When $\delta = 0.6$ the four subclusters are visible and samples are distributed according to class labels in different regions of the space considering simultaneously discriminative information and latent structure of the data. Finally when $\delta = 1$ the K_δ kernel matrix only keeps the structure of the data following the ground truth distribution but it does not retain any information about the classes.

3.6 Datasets

In this section the feature selection methods described previously are evaluated on synthetic data and on real world data using three tumor datasets.

3.6.1 Synthetic dataset

The goal of using a synthetic dataset is to know beforehand which features are informative and which ones are noise as ground truth. The feature selection methods are evaluated for the accuracy in selecting informative features as explained in section 3.4.6. The synthetic dataset has been generated by the *make-classification* function of Scikit Learn library [63] with two classes, $n = 200$ samples and $d = 200$ features. The dataset has been designed to be difficult to process since the informative features are 5% of the total number of features and the rest are considered noisy features. Additionally each class has been composed by 15 subclusters to make the data structure an important aspect of the problem. Therefore the discriminative information between each class and the data structure composed by class sub-clusters are determined by the 10 informative features.

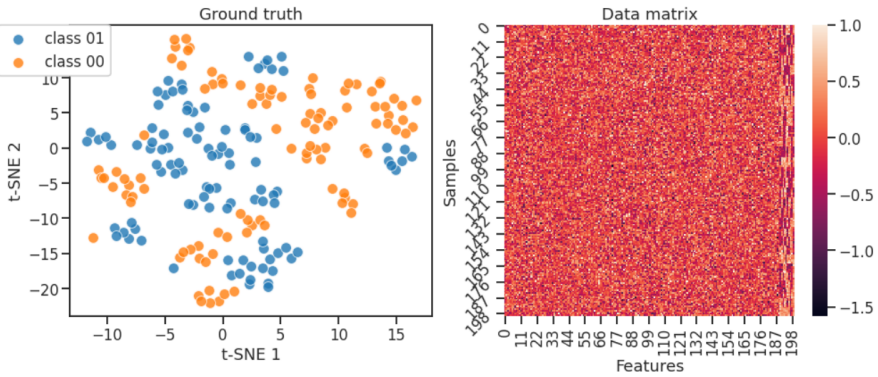

Figure 3.18: Ground truth distribution (left) and data matrix (right) of the synthetic dataset.

Figure 3.18 shows a two-dimensional scatter plot of the ground truth distribution of the synthetic dataset where class subclusters are visible and the data matrix where informative features are observed in the right columns.

3.6.2 Real world datasets

Three cancer datasets containing tumor profiles characterized by gene expression (RNA-Seq) features are used to evaluate the proposed method. These datasets are the Breast Cancer BRCA-US, the Pancreas Cancer PACA-CA and the Lung Cancer SMK-CAN-187. The BRCA-US and PACA-CA datasets are available from the International Cancer Genome Consortium [16]. The SMK-CAN-187 has been presented by [64] and available from `http://featureselection.asu.edu/` and the Gene Expression Omnibus [65] under GEO

accession number GSE4115.
The Breast cancer BRCA-US dataset is composed by 194 Breast cancer samples labeled by the survival days since diagnosis. To define two classes the threshold between low and high survival is defined as five years of survival since diagnosis [66]. The BRCA-US tumor samples are characterized by the expression of 20502 protein coding genes.
The Pancreas cancer PACA-CA dataset is composed by 135 tumor profiles labeled by tumor stage. The stage labels are IA, IB, IIA and IIB. The stages IA and IB are considered early stage while the stages IIA and IIB are considered late stage. The PACA-CA tumor samples are characterized by the expression of 18020 protein coding genes.
The SMK-CAN-187 is a benchmark microarray based gene expression database of Lung Cancer and it has 187 samples and 19993 gene expression features.

Dataset	Gene features (d)	Class (y)	Samples (n)
BRCA-US	20502	Low Survival	63
		High Survival	131
PACA-CA	18020	Early Stage	84
		Late Stage	51
SMK-CAN-187	19993	Control	90
		Tumor	97

Table 3.3: Size of each dataset.

Table 3.3 summarizes the size of each data set. It is clearly visible the high dimensional context of the three datasets and the low number of tumor samples.

3.7 Experimental results

In this section experimental results on the synthetic and real tumor data are reported.

3.7.1 Experiments on synthetic data

In this section it is evaluated the performance of each method in selecting the informative features by using the Feature Selection Accuracy score. The feature selection task is evaluated 50 times by randomly splitting 80% of train and the remaining 20% as test data. Then accuracy results are averaged across all iterations.

Latent regularization	FS_{acc}
$\delta = 0$	0.32 ± 0.01
$\delta = 0.2$	0.32 ± 0.01
$\delta = 0.4$	**0.33 ± 0.01**
$\delta = 0.6$	0.31 ± 0.01
$\delta = 0.8$	0.25 ± 0.01
$\delta = 1$	0.12 ± 0.01

Table 3.4: Feature selection accuracy for different values of δ parameter in the Kernel Latent Regularization step.

Table 3.4 shows how the Latent Regularization slightly improves the feature selection accuracy when $\delta = 0.4$. When compared with the benchmark methods in table 3.5 the proposed KLR-FS outperforms all the baselines and achieves the highest feature selection accuracy followed by RFE and HSIC-Lasso. The lower accuracy is achieved by the mRMR.

Method	FS_{acc}
KLR-FS ($\delta = 0.6$)	**0.33 ± 0.01**
HSIC-Lasso	0.27 ± 0.02
mRMR	0.20 ± 0.02
RFE	0.28 ± 0.02

Table 3.5: Feature selection accuracy for proposed KLR-FS and benchmark methods.

3.7.2 Experiments on real data

The initial sample set has been randomly split between 80% as train and 20% as test. By using just the training samples all the gene expression features have been auto-scaled with 0 mean and unit variance. Test samples have been scaled using the transformation learned from train samples. Then by using training samples the feature selection methods are applied and a subset of features are selected. With the selected features a support vector classifier is trained and tuned by 5-fold cross validation within train set for hyperparameter selection and evaluated on the test set.
For performance estimation the split between train and test set has been repeated randomly five times and each time a feature selection and classification tasks have been implemented. Then the classification results and the redundancy rate of the feature selection are averaged across all random splits and the mean and standard deviation of these metrics are reported. The experimental results section is divided in two subsections. The first one is devoted to explore the Latent Regularization for feature selection method and analyzes how the performance of KLR-FS behaves for different values of the mixture coefficient δ described in

equation 48. The second subsection is the performance estimation and statistical comparison of the proposed method and the benchmark ones.

3.7.3 Latent regularization for feature selection

This subsection details how the feature selection process of KLR-FS behaves with different settings of the mixture coefficient δ. As explained in the materials and methods section by varying the values of δ different target kernels k_δ are obtained and their corresponding Gram Matrix built, each one with a different mix between the unsupervised k_z and the supervised k_{yy} kernels. A set of six values of the Mixture Coefficient $\delta = [0, 0.2, 0.4, 0.6, 0.8, 1.0]$ are evaluated by decreasing the dependency to the labels. Each value of δ determines a feature selection model. Then the resulting kernel k_μ is used for support vector classification and evaluated on the test set. Figure 3.19 shows the AUC results on classification by using features selected for different values of δ on the three datasets. The AUC score peaks between $\delta = 0.4$ and $\delta = 0.6$ in all the datasets and number of selected features p. This results evidence the importance of the latent regularization in classification and means that the maximum AUC score using the KLR-FS features is obtained with a mixture between supervised labels and unsupervised latent structure.

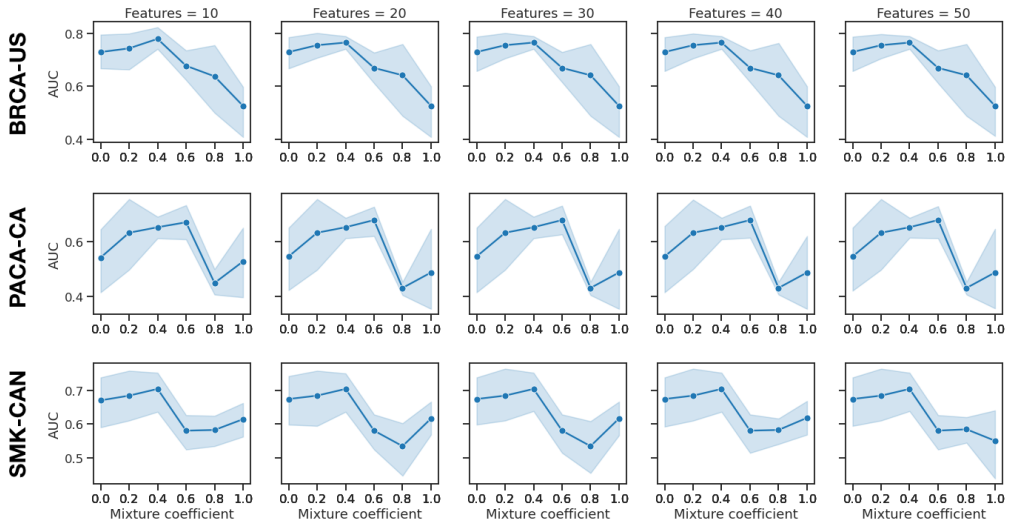

Figure 3.19: Classification performance measured by AUC-ROC using different number of selected features by KLR-FS across different values of the δ mixture coefficient.

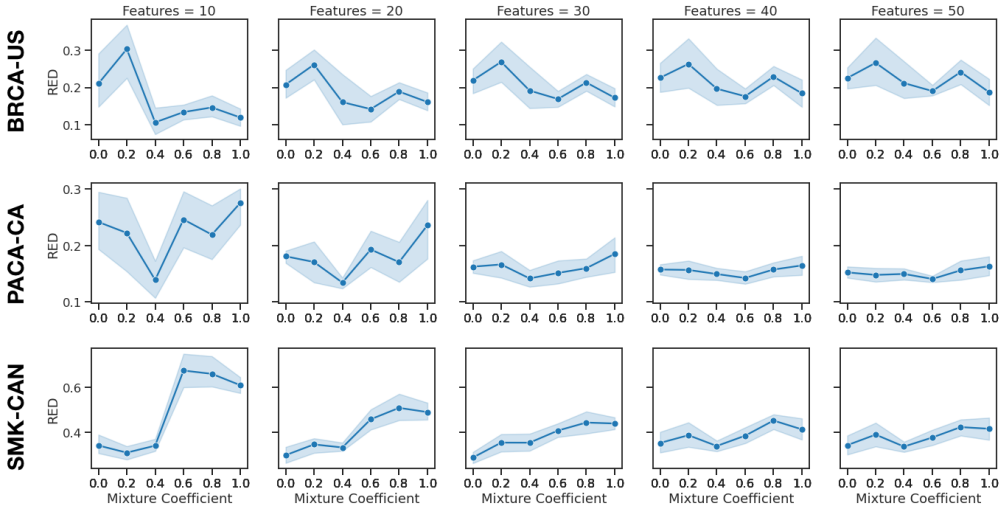

Figure 3.20: Redundancy rate RED of the selected features by KLR-FS with different values of the δ mixture coefficient.

Figure 3.20 shows how the RED score varies for different values of the Mixture Coefficient δ on the corresponding selected features and how the RED score is reduced by the latent regularization. For the BRCA-US and PACA-CA datasets the lowest values of RED are obtained for $\delta = 0.4$ and $\delta = 0.6$ and match with the same range of δ values with maximum AUC. For the SMK-CAN dataset the lowest RED scores are obtained between $\delta = 0$ and $\delta = 0.4$ and only when the number of selected features $p = 20$ and $p = 30$ the lowest RED score is at $\delta = 0$. Figure 3.20 shows that when the number of selected features p is small the mixture parameter δ has a stronger impact in the redundancy of the selected features.

3.7.4 Performance estimation and statistical comparison

In this subsection the KLR-FS is compared with the benchmark methods for different number of selected features p. For performance estimation the results of the classification and selection tasks are averaged among all the random iterations. The feature selection is applied on train samples and the classifier is trained on the selected features to finaly classify the test samples. The mixture coefficient used for KLR-FS for benchmark is $\delta = 0.4$ since it appear to be the best combination according to AUC and RED scores in the previous section.

Figure 3.21 shows the classification performance on test set for different numbers of p selected features by each method. For the BRCA-US and PACA-CA datasets the KLR-FS shows the highest mean AUC score and the lowest classification variance for every number of selected features. For the SMK-CAN dataset the KLR-FS has the highest classification performance with $p = 10$ features and for larger number of features it is outperformed by the RFE method.

In the three datasets the classification results using the KLR-FS features are the most constant

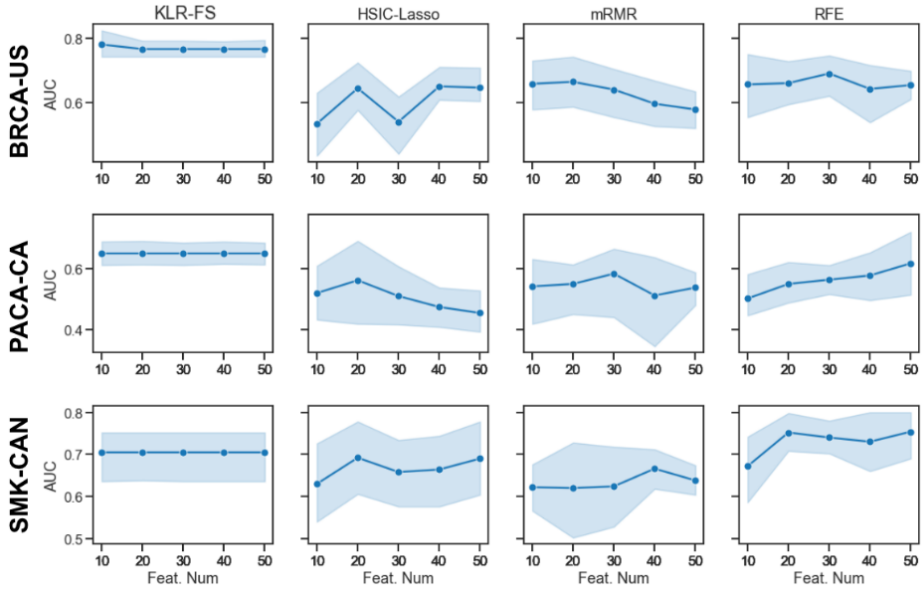

Figure 3.21: Comparison of the classification performance measured by AUC-ROC between KLR-FS and benchmark methods for different number of selected features p and different datasets.

as the value of p increases. According to section 3.3 the resulting k_μ assigns a larger weight μ_i to the variables that are selected first since these are the ones that increase more the overall alignment $A(K_\delta, K_\mu)$, while the variables selected later corresponds to smaller weights. The KLR-FS results on Figure 3.21 suggests that these classification problems can be solved by selecting the first 10 variables since selecting more variables does not introduce a significant improvement in the classification results.

Figure 3.22 shows the RED score of each subset of selected features for different values of p. For the BRCA-US and PACA-CA datasets the lowest score is obtained by the mRMR method with a $RED < 0.1$ followed by RFE, KLR-FS and HSIC-Lasso. For the SMK-CAN dataset the lowest RED score is obtained by the RFE method followed by HSIC-Lasso and KLR-FS while the highest corresponds to mRMR. KRL-FS and HSIC-Lasso shows a similar behaviour with a $RED < 0.2$ in the BRCA-US and PACA-CA datasets and a $RED < 0.4$ for the SMK-CAN datasets. Despite KLR-FS does not achieves the lowest RED score it is has a similar performance to HSIC-Lasso and in some cases to RFE and as observed in figure 3.22 KLR-FS achieves the highest AUC score.

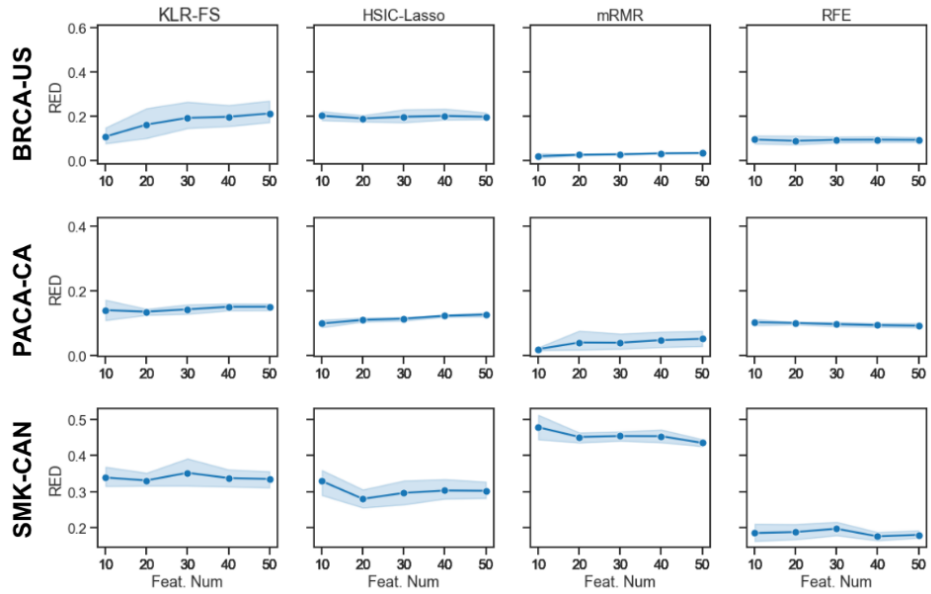

Figure 3.22: Evolution of the RED score for different number of features on each method.

This section compares the KLR-FS method with $\delta = 0.6$ and four feature selection methods. The main objective of this work is to select features to improve the tumor classification. It is observed that the classification performance of the KLR-FS is the highest for the majority of the cases.

3.8 Discussion

This work proposes a feature selection model based on Multiple Kernel Learning coupled with kernel-PCA with an application to tumor classification using high dimensional gene expression data from tumor profiles. The proposed method KLR-FS aims to select features considering not only the sample labels y but also the latent variables z of the training data presented in this work as a *latent regularization* since it is used to improve the generalization capacity in feature selection for classification. The idea is based on considering not only a pure supervised target kernel k_{yy} but also a kernel k_z built from the latent variables z obtained from a nonlinear dimensionality reduction ϕ_z. The latent regularization comes from a kernel built on the latent space $\mathcal{Z}$ generated by the kernel-PCA. This allows to explore the relaxation of the sample labels by a linear combination of a supervised and an unsupervised kernel. The experiments detailed on section 7 show that the highest classification performance and the lowest redundancy rate for the KLR-FS is obtained when latent information is mixed with label targeted information. These results show that selecting variables contemplating both the structure of the latent space $\mathcal{Z}$ of the data and the labels of the supervised problem can improve the classification task. Moreover, learning partially a latent space works as a regularization term since it limits the solution space by introducing the need to capture also

the general structure of the data. By doing so it can improve the generalization capacity on new unseen test samples. In addition, a relation between the latent structure of the data and the sample labels y exists and only learning from the labels with a classic supervised criteria can result in overfitting when training in a context where the number of samples is very limited. Therefore selecting features with a mix of supervised and unsupervised latent criteria may decrease the bias of a downstream classifier. Figure 3.19 reveals how the latent regularization works as a relaxation of the supervised problem by mixing K_{yy} with K_z. This evidences the important role of the unsupervised latent variables in the supervised feature selection task. As expected, in the unsupervised case with a value of $\delta = 0$ almost does not select any useful feature since the classification performance declines significantly in almost all cases. Finally, there is consistency between the peak in the AUC and the lowest RED score in the majority of the cases with a mixture coefficient ranging between $0.4 < \delta < 0.6$. From a Error perspective the Kernel Latent Regularization introduced in this work can be interpreted as a way to reduce the Approximation Error and the Bias Error since the classification performance peaks when it is used. Reducing the Bias error improves the generalization performance on new unseen test tumor samples.

Despite the kernel-PCA is used as a nonlinear and unsupervised mapping function $\phi_z(x)$ to learn a low dimensional latent space from training data any other nonlinear transformation could be used like Autoencoders or t-SNE [67] [68]. In this work since the tumor data has a considerably small sample set in comparison to the number of features $n << d$ then Autoencoders and t-SNE were not considered although they could be used in cases where more samples are available.

KLR-FS provides feature importance via the non-zero elements of the resulting sparse μ vector, which is a useful characteristic to gain interpretation of the results. The model shows that it can deal with high dimensional problems, in this case the initial dimension in all datasets is $d > 18.000$ and low sample sizes $n < 200$. In this high dimensional low sample context the classification using the KLR-FS features outperforms the benchmark methods in real data. Despite the KLR-FS is only outperformed in RED score by RFE in the three datasets and by mRMR in two datasets, KLR-FS has the highest classification performance and the lowest variance when compared with all the benchmark methods. Additionally in the synthetic dataset experiment the proposed KLR-FS method shows how the latent regularization term improves the feature selection accuracy and outperforms all the benchmark methods.

KLR-FS outputs a custom kernel k_μ that is used in support vector classification and contributes to improve the classification performance. The resulting k_μ kernel can be used not only for classification tasks but also for representation of tumor profiles or visualization by kernel-PCA since it is built from both supervised and unsupervised latent sources.

The proposed method has relevance in problems where the structure of the training data correlates partially with the tumor labels thus learning also from the latent structure of the data improves the supervised learning task.

3.9 Conclusion

In a context of low sample size and high dimensional space we propose a novel feature selection method that want to enforce the selected features to be discriminant and to keep information about the general data structure. By doing so we expect to reduce overfitting, to improve generalization and finally classification performance. To reach that goal we design a new approach called *Kernel Latent Regularization Feature Selection* (KLR-FS). It is based on a MKL approach that targets a label kernel relaxed with another kernel built on a latent space. In the proposed application the latent space is obtained using kPCA. The method is applied on high dimensional gene expression tumor profiles from Breast, Pancreas and Lung Cancer. KLR-FS selects genes which are used to classify tumor subtypes or survival rate with the highest classification performance and a considerably low redundancy rate when compared with benchmark methods. The mixture between supervised labels and latent variables reveals an improvement in the generalization capacity when compared with other feature selection methods.
This thesis chapter has been devoted to explore the supervised feature selection problem in a cancer genomics context where the number of features d overpass the number of samples n. The combination of kernel methods and latent variable models have shown improvements in the classification results. The next chapter explores an extension of the proposed kernel latent regularization term in Sufficient Dimension Reduction problems as a way to show that the proposed idea works in multiple approaches. Additionally the MKL feature selection method proposed can easily be extended to unsupervised feature selection problems as explained in chapter 05 and extended to multi-omic latent space in chapter 06 where genomics, transcriptomics, proteomics and metabolomics could be available and select features in a multi-modal approach.

3.9.1 Scientific Production of this chapter

The content of this chapter has been presented in:

- Preprint article *Latent regularization for feature selection using kernel methods in tumor classification* [31]
- Congress Paper at AGRANDA 2018 *Learning Kernels from genetic profiles to discriminate tumor subtypes* [69]
- Journal Paper *Coupled Mass-Spectrometry-Based Lipidomics Machine Learning Approach for Early Detection of Clear Cell Renal Cell Carcinoma* [70]
- SADIO Electronic Journal 2019 *Hepatocellular Carcinoma tumor stage classification and gene selection using machine learning models* [71]
- ICML 2020 LatinX in AI Workshop *Kernel Latent regularization for feature selection in tumor classification* [72]

Chapter 4

Survival analysis with Sufficient Dimensionality Reduction

4.1 Introduction

As explained in the previous chapters, dimensionality reduction plays a key role for tumor profile analysis. Chapter 02 detailed how a single tumor can be characterized by tens of thousands of gene expression features and the importance of dimensionality reduction. Chapter 03 studied supervised [31] feature selection to reduce dimensionality by keeping the interpretability of the reduction and to perform a downstream learning task like classification. Particularly, chapter 03 proposes the kernel latent regularization feature selection (KLR-FS) method [31]. The latent regularization term in KLR-FS aims to mix supervised and unsupervised kernel matrices to build a hybrid target kernel that guides the feature selection process. Therefore the selected features are the ones that captures information about the labels but also the latent structure $\mathcal{Z}$ of the training data. This latent regularization shows how the downstream learning task like classification is improved when the feature selection process considers the tumor labels and the unsupervised latent structure simultaneously. In this chapter the generalization of the concept of latent regularization beyond the KLR-FS method for feature selection is explored. One of the characteristics of the kernel latent regularization relies in the fact that it can be used on any supervised learning task that has a target kernel function during the learning process. For this reasons this chapter aims to expand the latent regularization concept to dimension reduction and feature extraction methods that uses Kernels as target functions. In this chapter the Sufficient Dimension Reduction method, a feature extraction approach, is studied instead of studying feature selection as chapter 03 does. The idea is to analyze how the Kernel Latent Regularization term can be used to improve a Sufficient Dimension Reduction task using a hybrid target kernel to learn a low dimensional representation of the input data.
Dimensionality reduction by feature extraction is commonly performed by Unsupervised learning approaches and is a task that can be performed by PCA [73], kernel PCA [74], Autoencoders [75], t-SNE [76] among others. These methods learn a low dimensional latent

space $\mathcal{Z}$ and have the objective function designed to optimize structure preservation of an input distribution X of interest. On the other side supervised learning approaches for dimensionality reduction and feature extraction are focused on information preservation about a target response y [77]. Particularly in this chapter the Sufficient Reduction (SDR) approach is studied [78]. The SDR involves the family of supervised dimensionality reduction methods used for classification and regression. As a supervised approach, SDR considers the sample labels y to learn a projection $\rho(x)$ of the input data $X \in \mathbb{R}^d$ in a low dimensional space $\mathcal{S} \in \mathbb{R}^p$ where $p < d$ such as [79]

$$\rho(\mathcal{X}) = \mathcal{S} \tag{4.1}$$

Formally, the subspace $\mathcal{S}$ is a *dimension-reduction subspace* in terms of the following conditional independence

$$Y \perp X \mid \beta X \tag{4.2}$$

where $\perp$ indicates independence and βX represents the orthogonal projection of X into $\mathcal{S}$.

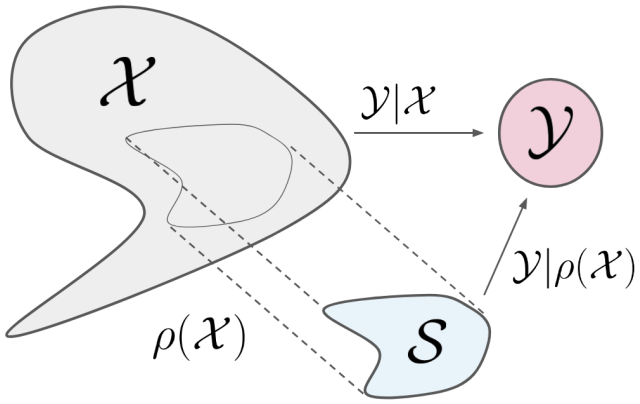

Figure 4.1: SDR concept.

Figure 4.1 shows a diagram of how the SDR method is designed. The SDR family of methods may be useful to predict tumor labels like prognosis [80]. SDR methods can be used to learn a subspace $\mathcal{S}$ that captures sufficient information about the relationship between the gene expression features and the phenotype outcome such as patient survival [81].
An important characteristic about SDR methods is that they only consider Y to estimate the subspace $\mathcal{S}$ and the unsupervised structure of the data is not used. This scenario may cause overfitting in a context where the number of gene expression features d is higher than the number of tumor samples n. The hypothesis studied relies in the idea that supervised learning tasks like dimensionality reduction can be improved by learning simultaneously from the sample labels and also from the latent structure of the input data. This chapter aims to answer the question: it is possible to improve a regression learning task coupled with Sufficient Dimensionality Reductions by introducing the Latent Regularization?
Since SDR approaches are designed for classification and regression in this chapter it is

studied how SDR can be used for patient survival analysis by learning a supervised low dimensional space. The proposal of this chapter is a method that improves the survival regression performed on the SDR projected space $\mathcal{S}$ by using a Kernel Latent Regularization method that learns from both tumor labels and the unsupervised latent structure of the training data [31].

4.2 Sufficient Dimensionality Reduction

Supervised dimensionality reduction methods are focused on information preservation about a target response y. In this line, this chapter is based on *sufficient* reductions. In its most common form, sufficient dimension reduction (SDR) is a methodology that deals with supervised learning problems involving a set of input features $X \in R^d$ and a dependent variable or response Y. SDR methods combines both dimension reduction and the sufficiency concept to find a low dimensional subspace $\rho(x) = \mathcal{S}_{Y|X}$ that captures the dependency of the target labels Y on X [82]. The learning task is to find a transformation of the input features $\rho(X) \in R^p$, with $p < d$, that encloses all the information available in X about Y. Formally it is said that ρ(X) is a SDR for $Y \mid$ X iff one of the following conditions are met [83]:

$$\text{i)} \quad X \mid (Y \mid \rho(X)) \sim X \mid \rho(X), \tag{4.3}$$

$$\text{ii)} \quad Y \mid X \sim Y \mid \rho(X), \tag{4.4}$$

$$\text{iii)} \quad X \perp Y \mid \rho(X) \tag{4.5}$$

where $\sim$ means identically distributed. These conditions entail the idea about the reduction $\rho(x)$ containing all the information of Y from X which is all the information used to estimate $E(Y|X) = E(Y|\rho(X))$. Most of the available methods [84] search for a linear transformation

$$\rho(X) = \beta^T X \tag{4.6}$$

where β is a $d \times p$ matrix, the actual goal in estimation is to determine β.One common approach for estimation is Sliced Inverse Regression (SIR) [77], a method centered on studying the conditional distribution $P(X|Y)$. The main idea of the inverse regression concept lies in the assumption that the conditional distribution $P(X|Y)$ is concentrated in a subspace $\mathcal{S}$ of the input space $\mathcal{X}$. Therefore the inverse regression $E(X|Y)$ may lie in the same subspace $\mathcal{S}$ [82].

Most of the inverse regression methods make a linear assumption with respect to the probability distribution of X when computing the regression $X \mid Y$. The linear assumption is a limitation that may constraint the possible distributions of X. Despite it is always possible to find a linear transformation that preserves information, it is not possible to guarantee that the characteristic information lies in a low-dimensional linear subspace $\mathcal{S}$ of the features $\mathcal{X}$. Additionally, if the subspace $\mathcal{S}$ required to preserve information is too large, the potential benefits of dimensionality reduction vanish.

Despite the standard SDR preserves discriminant information with respect to Y, the intrinsic structure of the data is not preserved and a multimodal structure of the input distributions may be lost. This is an issue when $\mathcal{S}$ is estimated based on a limited amount of data. Such as for cancer applications where the preservation of the intrinsic geometry may produce a more conservative SDR that enables to detect cluster subtypes within a tumor class, to perform further data analysis tasks and to improve generalization on unseen test data.

Example

Suppose a synthetic dataset (X, Y) with $n = 300$ samples characterized by $d = 100$ features as $\mathcal{X} \in \mathbb{R}^d$ and $y \in \{0, 1, 2\}$ a three classes label vector. The intrinsic dimensionality of the dataset is $p = 2$ and the $d = 100$ input dimension correspond to redundant features obtained from the intrinsic ones. The data of each class is subdivided in two subclusters, therefore the structure of the input data is composed by six clusters in total. A SDR reduction is applied and a p-dimensional subspace $\mathcal{S}$ obtained with $p = 2$ to visualize in low dimension the distribution of the data.

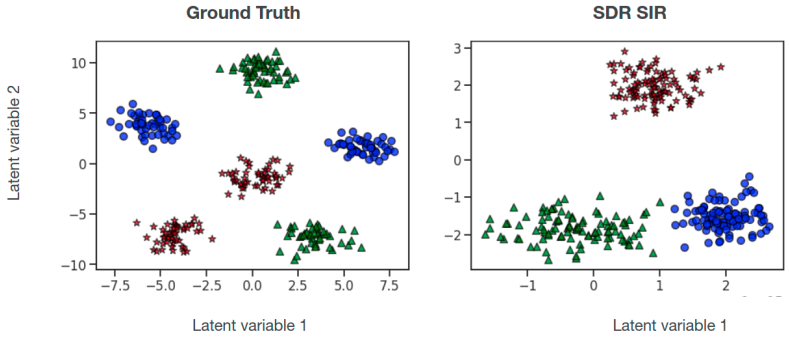

Figure 4.2: Supervised dimensional reduction via Sufficient Dimension Reduction (SDR) with SIR method. Left figure: Latent space of the ground truth distribution. Right: samples distributed in the resulting representation $\mathcal{S}$ by SIR.

Figure 2 shows a 2D scatter plot of the ground truth latent space of a synthetic dataset with three classes, each one composed by two sub-clusters. The latent characteristic subspace $\mathcal{Z}$ is visualized on the left panel. The right panel shows how the SIR SDR projects the input data X on the obtained subspace $\mathcal{S}$. It successfully captures the conditional distribution between X and Y and the discriminant information is preserved. Nevertheless the intrinsic manifold of the data is completely lost and the class subclusters are not preserved. Therefore, standard SDR methods fails in keeping the structure of the data causing limitations in downstream analysis on the obtained representation. Additionally, when the database is limited missing out the structure of the data may cause overfitting and a limitation in the generalization capacity of the learned subspace $\mathcal{S}$ for supervised tasks like classification or regression.

Following this motivation example this chapter proposes a further analysis in methods that keep the intrinsic (latent) structure of the input data and the discriminant information about a target response Y simultaneously via sufficient reductions on gene expression data from tumor profiles.

4.3 Kernel-Dimensionality Reduction

The SIR method for SDR reduction presented in the previous section fails in large d and small n settings due to the operational complexity that is faced when computing the covariance matrix for the estimation of β [85]. Therefore the SIR methods are not suitable for SDR in a cancer genomics context where $n < d$. A different strategy is to exploit properties of the input features $\mathcal{X}$ in a reproducing kernel Hilbert spaces (RKHS) [86] in order to characterize the conditional independence

$$Y \perp X \mid \rho(x) \tag{4.7}$$

known as *kernel sufficient dimension reduction* (kSDR) [87]. Let $k_X : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $k_y = \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be two positive definite kernels inducing two RKHS $\mathcal{F}$ and $\mathcal{G}$ respectively. There exists a conditional covariance operator $\mathbf{\Sigma_{Y|X}} : \mathcal{G} \to \mathcal{G}$ for any function $g \in \mathcal{G}$ so that

$$\langle g, \mathbf{\Sigma_{Y|X}}\, g\rangle_{\mathcal{G}} = \mathbb{E}[\mathrm{Var}_{\mathbf{Y}|\mathbf{X}}(g(\mathbf{Y})|\mathbf{X})].$$

where the covariance operator computes the residual error of predicting $g(\mathbf{Y})$ with X. An analog definition applies for $\Sigma_{\mathbf{Y}|\mathbf{B}^T\mathbf{X}}$ when using a linear reduction $\mathbf{B}^T\mathbf{X}$ as predictor. The decisive property of such covariance operators is that, for any projection matrix $\mathbf{B}$, $\mathbf{\Sigma_{Y|B^TX}} \geq \mathbf{\Sigma_{Y|X}}$ in the partial order of the trace operator, with equality only when $F(\mathbf{Y}|\mathbf{X}) = F(\mathbf{Y}|\mathbf{B}^T\mathbf{X})$. Thus, by minimizing an estimate of $\mathrm{Tr}(\mathbf{\Sigma_{Y|B^TX}})$ as a function of $\mathbf{B}$ we can minimize the loss of predictive information when replacing $\mathbf{X}$ with $\mathbf{B}^T\mathbf{X}$ [88]. The working criterion is to minimize over $\mathbf{B}$ the kernel-based measure of independence

$$\mathcal{J}_{YY|X}(B^\top X, Y) = \mathrm{Tr}(\bar{\mathbf{K}}_{\mathbf{Y}}(\bar{\mathbf{K}}^B_{\mathbf{X}} + n\epsilon_n\mathbf{I})^{-1}), \tag{4.8}$$

where $\bar{\mathbf{K}} = \mathbf{CKC}$, $\mathbf{C}$ the centering matrix given by $\mathbf{C} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^T$, $\mathbf{K_Y}$ and $\mathbf{K}^B_{\mathbf{X}}$ and kernel matrices corresponding to Y and $\mathbf{B}^T X$ respectively, ϵ is a regularizer that smooths the kernel matrix.

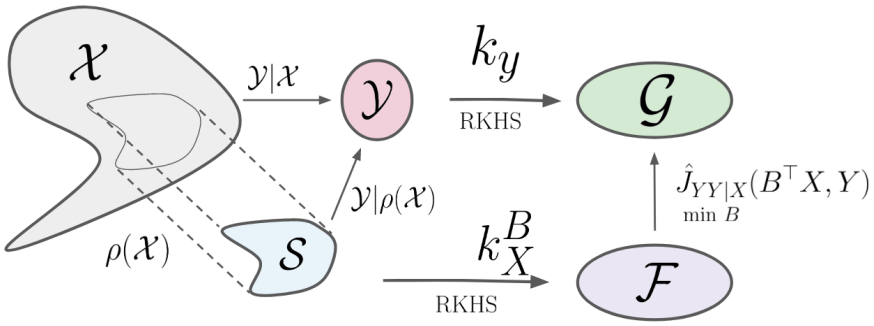

Figure 4.3: Diagram of the kSDR method.

Figure 4.3 shows a diagram of the kSDR approach and how the kernel methods are used to learn the low dimensional representation $\mathcal{S}$.

Example

Continuing with the example of the synthetic dataset presented in the previous section the kSDR method is applied to obtain a $\mathcal{S}$ reduced subspace. Then $\mathcal{S}$ is compared with the ground truth latent distribution of the data.

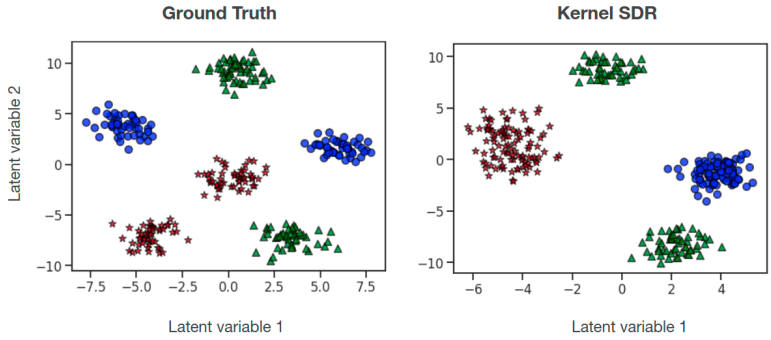

Figure 4.4: Left: Ground truth latent distribution of the training data. Right: obtained subspace after kSDR reduction.

Figure 4.4 shows on the left figure the ground truth latent distribution of the input data. The data samples are labeled by three classes, each one structured with two sub-clusters. The figure on the right shows the obtained subspace $\mathcal{S}$ after applying kSDR. It is observed that kSDR captures partially the sub-clusters of the training data but the ground truth latent structure is not completely preserved. Since kSDR is only focused on Y the resulting subspace makes no difference between the subgroups of each class.

4.4 Proposed method: Latent regularization Kernel Sufficient Reduction

Section 3 shows how SDR in both SIR and kSDR methods may fail in capturing simultaneously the label information y and the structure of the data when reducing dimensionality. Additionally, if the SDR task only keeps label information y it can be exposed to overfiting when dealing with high dimensional and noisy data therefore may affect the generalization to new unseen data. This work proposes a method to solve both problems: structure preservation and improvement of generalization capacity in situations when the number of samples n is limited compared to the dimension d.
As explained in the previous section, the kSDR can be performed by minimizing the kernel-based measure of independence $\mathcal{J}_{YY|X}(B^\top X, Y)$. This measure is computed by using the k_y and k_x kernels built on Y and X respectively. Therefore it is possible to consider k_y as a *target* kernel used to obtain the reduced subspace $\mathcal{S}$. The original kSDR method [82] considers only a supervised kernel k_y built with the sample labels y as target. If additional information such as latent structure Z of the training data is available then it can be stored as an unsupervised kernel K_z and included in the target kernel as well. This work proposes to use the kernel latent regularization concept introduced in the chapter 03 of this thesis [31] to make the target kernel a hybrid one K_δ composed by supervised information of the labels but also by unsupervised latent structure of the data such as

$$\mathbf{K}_\delta = (1-\delta)\mathbf{K}_y + \delta\mathbf{K}_z \tag{4.9}$$

where $\delta \in (0,1)$ is the *mixture coefficient* used to balance the contribution between a supervised kernel K_y built with tumor labels and an unsupervised kernel K_z built on the latent space $\mathcal{Z}$. The latent space can be obtained from a mapping function ϕ such as $z = \phi(x)$ with $z \in \mathbb{R}^p$. It is assumed that the high dimensional input space $\mathcal{X}$ does not reveals clearly the structure of the data due to noisy and redundant features therefore an unsupervised dimension reduction method $\phi(x)$ is used to get the low dimensional latent space and capture the data structure. In this work the mapping $\phi(x) = z$ is learned via the kernel-PCA method [74] nevertheless any other dimensionality reduction method such as t-SNE [76], Autoencoders [75], PCA [73] and UMAP [89] can be used to learn ϕ. Then the kernel based measure of independence at equation (8) can be re-written as

$$\mathcal{J}_{(Y,Z)|X}(B^\top X, (Y,Z)) = \mathrm{Tr}(\bar{\mathbf{K}}_\delta(\bar{\mathbf{K}}^B_{\mathbf{X}} + n\epsilon_n\mathbf{I})^{-1}), \tag{4.10}$$

By this simple replacement of the original target kernel K_y to K_δ it is possible to learn a low dimensional sufficient space $\mathcal{S}$ that captures the dependency between Y and X but also with Z which keeps the latent structure of the data. The proposed kernel latent regularization approach computes an estimate between the standard purely supervised case ($\delta = 0$) and the fully unsupervised scenario ($\delta = 1$).
The hybrid target kernel is proposed as a solution to the problem of structure preservation in supervised SDR methods as shown in figure 2 and 4 and to improve the generalization

capacity assuming that unseen test samples will follow a similar latent structure. Then instead of learning a reduction $\rho(x)$ that keeps the conditional distribution

$$y \mid \rho(x)$$

this work proposes to learn a SDR reduction $\rho(x)$ conditioned by tumor labels and latent structure simultaneously as

$$(y, z) \mid \rho(x) \to (y, \phi(x)) \mid \rho(x), \tag{4.11}$$

The proposed method is named Kernel Latent Regularization Sufficient Dimension Reduction (KLR-SDR).

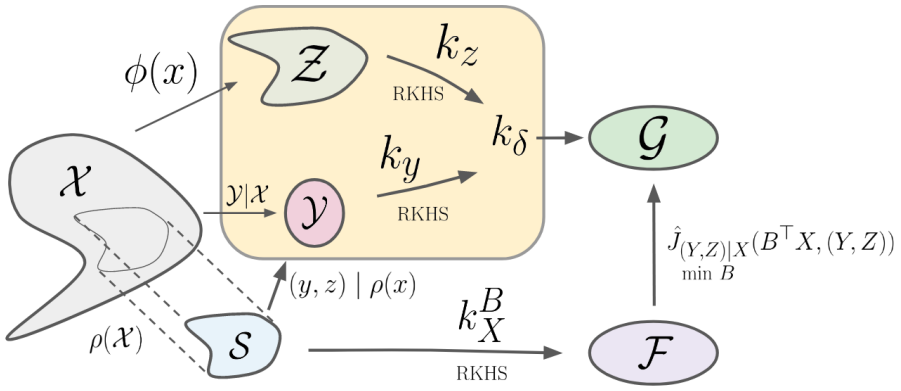

Figure 4.5: Proposed KLR-SDR method. The hybrid kernel k_δ mix supervised and unsupervised information from k_y and k_z respectively. Then the obtained reduction subspace $\mathcal{S}$ explains both labels and latent structure of the traning data.

Figure 4.5 shows the diagram of the proposed method KLR-SDR. The kSDR standard method is coupled by an unsupervised dimension reduction task that captures a latent space $\mathcal{Z}$ from the input space $\mathcal{X}$. An unsupervised k_z kernel is built on the latent space to capture the intrinsic structure of the data. Then k_z kernel is mixed with the supervised kernel k_y which contains the label information of the tumor samples to build a k_δ kernel that contains supervised and unsupervised information of the training data. Finally the reduced subspace $\mathcal{S}$ is computed by using k_δ as target kernel. Therefore, the obtained subspace $\mathcal{S}$ will be a sufficient dimension reduction of both the label and structure information.
The idea of keeping the structure accounts for the *ignorance* about the initial dataset information content and also to the data density that is so low in the initial space that smaller clusters cannot be found. This idea also may improve the generalization capacity of the learned representation on new unseen test samples.

Example

Let the synthetic dataset (X,Y) with three classes and two subclusters per class from the example of section 3. To apply the KLR-SDR first an unsupervised latent space Z of the input data is obtained via a non-linear dimnsion reduction like kernel-PCA as $z = \phi(x)$. Then a gaussian kernel is built on Z and a K_z gram matrix obtained. Then by multiple values of the mixture parameter δ different target kernel matrices K_δ are built as shown in figure 4.6 where gram matrices are visualized for different mixtures between unsupervised K_z and supervised K_y kernels. The higher the mixture coefficient δ the stronger the influence of the latent structure of the training data.

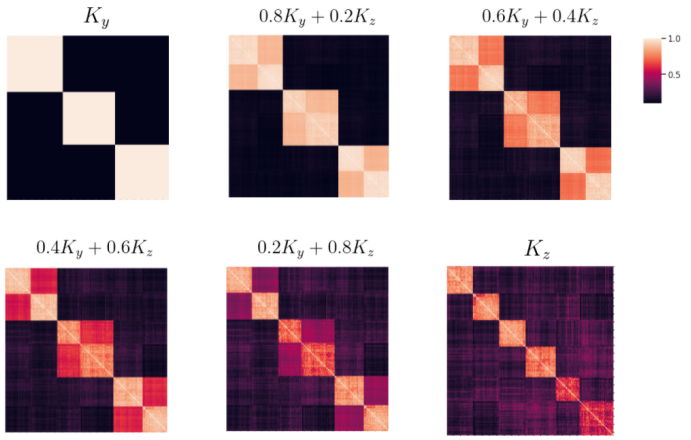

Figure 4.6: Gram matrices for different values of the mixture parameter δ. The upper left matrix is the supervised one where $\delta = 0$. The lower right matrix is the unsupervised one where $\delta = 1$. The rest of the matrices corresponds to values of the mixture parameter $0 < \delta < 1$.

With the mixture parameter $\delta = 0.5$ a reduced subspace $\mathcal{S}$ is obtained. The KLR-SDR representation is compared with standard SIR and kernel-SDR reductions as shown in figure 4.7. Additionally the ground truth latent structure of the data is shown in the upper left plot.

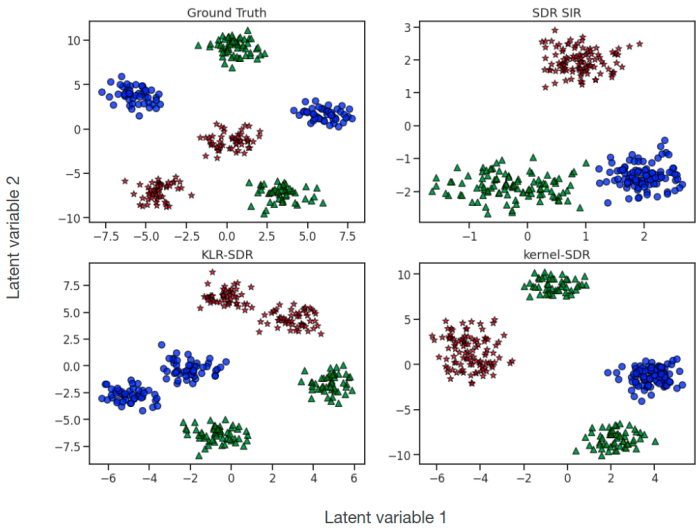

Figure 4.7: Two dimensional visualization of the resulting subspace $\mathcal{S}$ by each sufficient dimension reduction method: SDR SIR, Kernel SDR and the proposed KLR-SDR.

Figure 4.7 shows in the lower left picture how the KLR-SDR method keeps the discriminant information and simultaneously preserves the subclusters of each class when $\delta = 0.5$. On the other side, the SIR (upper right) and standard kSDR (lower right) methods fail to capture both structure and supervised labels simultaneously.

4.5 Experiments

The proposed KLR-KSDR method is evaluated with synthetic and real cancer data. The objective of the experimental section is to show how a supervised task like regression or classification trained on the subspace $\mathcal{S}$ is improved when the dimension reduction is affected by the latent regularization for different values of $\delta = [0, 0.2, ..., 0.8, 1]$. The KLR-KSDR method is compared with the standard kSDR, PCA and kernel-PCA as benchmark methods.

4.5.1 Datasets

The synthetic dataset is two-class problem composed of $n = 200$ and $d = 200$ features, where only 20 features are informative and the rest are noisy and non useful features. The synthetic dataset has been generated by the *make-classification* function of Scikit Learn library [63] where the discriminative information between each class composed by 3 sub-clusters each one is determined by the 20 informative features. Since the data structure is defined in high dimension and both classes are highly overlapped the three subclusters per class are not clearly visible when projected in two dimensions for visualization as presented in Figure 8. The labels defines a binary classification problem as $y \in \{-1, 1\}$. Additionally, to make the dataset difficult to classify with a standard supervised approach both classes are highly overlapped

with the objective to show how the proposed Kernel Latent Regularization approach can deal in these complex cases. Table 1 shows the size of the synthetic dataset.

Classes	samples (n)	features (d)	informative features	clusters per class
2	200	200	20	3

Table 4.1: Synthetic dataset for classification.

The real world data corresponds to four real datasets downloaded from the International Cancer Genome Consortium [90]. The real datasets are detailed in table 2

Type	Subtype	Samples (n)	Gene features (d)
Lung	LUAD-US	118	17233
Breast	BRCA-US	100	17233

Table 4.2: Cancer gene expression datasets for survival regression.

Each dataset corresponds a tumor subtype. Each patient sample in each dataset is labeled by its survival time T from positive diagnosis.

Given a gene expression tumor profile x characterized by d expressed genes the goal is to learn a regression function trained on the reduced subspace $\mathcal{S}$ of dimension p with $p < d$ to estimate the survival rate of patients.

4.5.2 Evaluation

To evaluate the quality of the proposed method a downstream supervised learning task is used on the obtained subspace $\mathcal{S}$ for both the synthetic and real datasets.
For the synthetic dataset a support vector classification task [62] is performed to classify samples x based on their binary labels y as

$$\hat{y}(\mathbf{x}) = \sum_{i=1}^{M} \alpha_i y_i \left\langle \phi_\mu(x), \phi_\mu(x_i) \right\rangle_{\mathcal{H}} + b = \sum_{i=1}^{M} \alpha_i y_i k_\mu(\mathbf{x}, \mathbf{x}_i) + b$$

and the evaluation is done by the Area Under the Receiving Operation Characteristic Curve (AUC) [36].
For the real world survival data from cancer patients a survival function named *Hazard Rate*

$$H(t \mid x) = P(T > t) \tag{4.12}$$

is estimated as a function of t and input covariates x. The $H(t \mid x)$ functions returns the probability of survival beyond time for each cancer patient. To estimate the $H(t \mid x)$ function the Cox Regression model is used [91] which lets identify the patients which are still alive given a time t. The Cox Regression model is based on the idea that the log-hazard rate of a sample is a linear function of their features and a baseline hazard function expressed as

$$H(t \mid x) = \exp(x^\top \beta) H_0(t) \tag{4.13}$$

where β are the regression coefficients to be estimated and $H_0(t)$ the baseline hazard function which is estimated by the Breslow estimator [92].
To evaluate the performance of the Cox Regression model the concordance index (C index) is computed [93]. The analytical expression [94] of the c-index is expressed as

$$C_{\text{index}} = \frac{\sum_{ij} 1_{T_j<T_i} 1_{H_j<H_i}}{\sum_{ij} 1_{T_j<T_i}} \tag{4.14}$$

where $1_{T_j<T_i} = 1$ if $T_j < T_i$ and else 0, $1_{H_j<H_i} = 1$ if $H_j < H_i$ else 0. When dealing with survival models the C index is one of the most frequently used evaluation metrics. It is a measure that ranges between 0.5 and 1 and it is a rank correlation between estimated risk scores $\hat{H}$ and observed survival labels T. Given a time t it is desired that the survival model assign higher risk for patients with lower survival time and viceversa. The C-index is an extension of the ROC curve, therefore a C-index close to 1 means a model that estimates accurately the survival time of a patient while a C index close to 0.5 does not. The C-Index is used to analyze how well given a time t a predictive model classifies samples between the ones that experience decease by time t (sensitivity) from those who will not (specificity) [95].

4.5.3 Experiments on Synthetic data

The synthetic dataset detailed in the previous section is visualized by a scatter plot on the first two component of a t-SNE embedding in Figure 4.8. It can be observed that both classes are highly overlapped and there is not enough separability between classes composed by three subclusters each one. Additionally the data has 20 informative features (Table 1) thus visualization in two dimensions may not be fully informative.

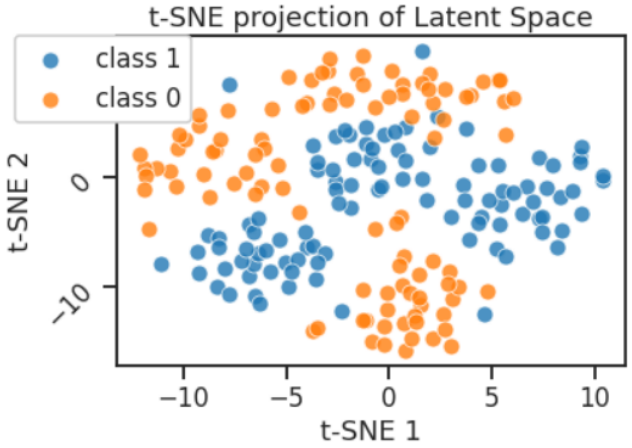

Figure 4.8: A 2 dimensional scatter plot using the projection in two axis of t-SNE method of the synthetic dataset.

To use the KLR-SDR method first a kernel-PCA model learns a latent space $\mathcal{Z}$ of the training data. The kernel-PCA model uses a gaussian kernel $k_{\text{rbf}}(x_i, x_j) = \exp\left(-\gamma \left\|\mathbf{x}_i, \mathbf{x}_j\right\|^2\right)$ and the γ parameter is computed as the median pairwise distance between samples. The latent space $\mathcal{Z}$ captures the inner structure of the data, therefore a gaussian kernel $k_{\mathcal{Z}}$ is built on the latent space and used as the unsupervised kernel.
On the other side, with the binary labels a supervised kernel K_y is built where $k_y(x_i, x_j) = 1$ if two samples belong to the same class and $k_y(x_i, x_j) = 0$ if samples are from different classes. Then following equation 9, by varying the mixture coefficient δ different mixed kernels K_δ are obtained as $\mathbf{K}_\delta = (1 - \delta)\mathbf{K}_y + \delta\mathbf{K}_z$.

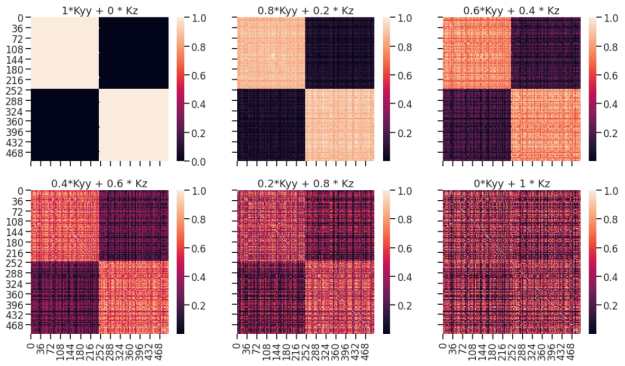

Figure 4.9: Gram matrices of K_δ for different values of the mixture parameter δ.

Figure 9 shows different kernel matrices of K_δ for $\delta = [0, 0.2, 0.4, 0.6, 0.8, 1]$. A kSDR dimension reduction model is trained for each K_δ, therefore the kSDR with $\delta = 0$ is a standard supervised kSDR while $\delta = 1$ is a completely unsupervised one. All the values of $0 < \delta < 1$ corresponds to different mixtures between supervised and unsupervised structure. In this experiment the dimension of the obtained reduction $\mathcal{S}$ is $p = [3, 5, 10]$. To estimate the classification performance of different values of δ the initial dataset has been splited 20 times between train and test set with a train set of 50% of the total samples. On each iteration a classifier is learned with train data and evaluated on the test set. The classification results are averaged among all unseen test set splits.
Figure 10 shows the average and standard deviation AUC ROC score of the SVM classifier trained on the obtained subspace $\mathcal{S}$ from kSDR for different values of δ. It is observed that the AUC ROC peaks at $\delta = 0.9$ evidencing the role of the proposed kernel latent regularization.

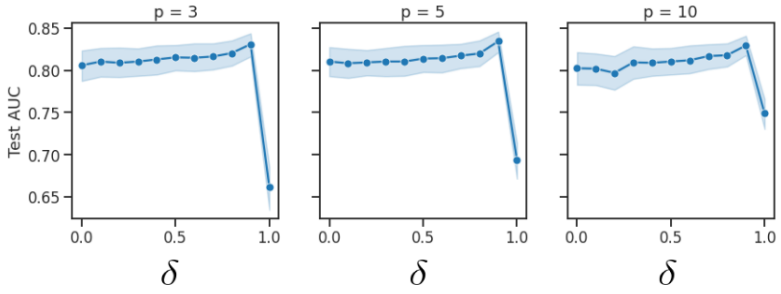

Figure 4.10: Mean Classification AUC-ROC for different values of δ with $p = 3$, $p = 5$ and $p = 10$. The peack performance is obtained with $\delta = 0.9$

The dimension reduction of the best regularized KLR-SDR model with $\delta = 0.9$ (figure 4.10) is compared using the classification results of a classifier model trained on the latent space obtained from a standard kSDR, PCA and kernel-PCA. Additionally the resulting baseline of learning a classifier directly from the input data ($d = 200$) is included in the analysis to prove the interest of the dimension reduction task.

Dim (p)	kSDR	KLR-SDR	PCA	kPCA	Baseline
3	0.80 ± 0.04	$\mathbf{0.83 \pm 0.03}$	0.55 ± 0.12	0.58 ± 0.11	0.74 ± 0.04
5	0.81 ± 0.03	$\mathbf{0.84 \pm 0.03}$	0.59 ± 0.11	0.58 ± 0.13	0.74 ± 0.04
10	0.81 ± 0.04	$\mathbf{0.83 \pm 0.02}$	0.65 ± 0.10	0.65 ± 0.08	0.74 ± 0.04

Table 4.3: Classification results in terms of p after dimension reduction on independent test set.

Results in Table 4.3 shows how the proposed KLR-SDR method with latent regularization outperforms the benchmark methods by learning a latent space $\mathcal{S}$ that improves classification generalization ability. The learned representation of KLR-SDR is even better than the input baseline data with $d = 200$ showing how the learned reduction keeps the discriminant and structural information in a lower number of dimensions.

4.5.4 kSDR for Survival prediction of cancer patients

The KLR-SDR method is also evaluated on real gene expression data from cancer patients for survival analysis. A separately experiment is carried out for each dataset. The dimension after reduction of the obtained subspace S is $p = 10$ for all the cases. Each dataset is randomly splited between train and test 20 times with a training set 70% of the total dataset.

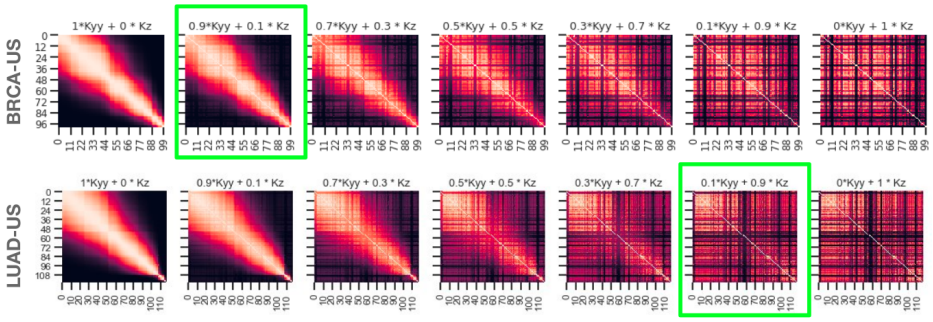

Figure 4.11: Kernel matrices K_δ for each tumor dataset for different values values of δ. On the left kernel matrices are completely supervised K_y. On the right kernel matrices are completely unsupervised K_z. Kernel matrices in the middle are the ones corresponding to different mixtures between K_y and K_z. Highlighted in green the kernel matrices associated with the highest survival prediction performance.

The unsupervised kernel matrix K_z is obtained with the same procedure as described in the synthetic data experiments. Nevertheless in this case the tumor labels are survival days as $y \in \mathbb{R}$, for this reason the supervised kernel K_y is built by applying a gaussian kernel on the labels and the γ parameter is the median pairwise distance between sample labels. Then the K_δ kernel containing the mix between supervised and unsupervised structure is built following equation 9. Figure 11 shows the resulting K_δ for $\delta = [0, 0.1, 0.3, 0.5, 0.7, 0.9, 1]$. With the K_δ kernel multiple reductions are performed using the proposed KLR-SDR method. For each value of δ a low dimensional reduced subspace S_δ is obtained and survival regression is performed.
Figure 4.12 shows the C-index of the survival estimation for each dataset across different values of δ. In the BRCA-US dataset the maximum C-index is obtained at $\delta = 0.1$ and for LUAD-US at $\delta = 0.9$. Figure 4.11 highlights in green the target kernel matrices associated to the KLR-SDR reduction with the highest C-index on each dataset. These results improved by kernel latent regularization are compared with the standard kSDR, PCA and kPCA.

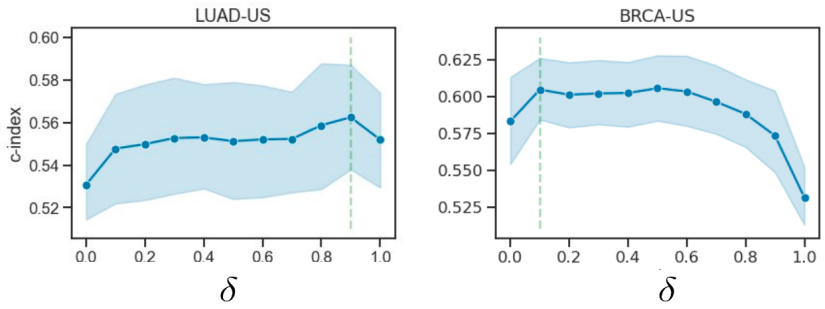

Figure 4.12: C-index of a Cox Regression model trained on the reduced subspace $\mathcal{S}$ from a KLR-SDR for different values of δ on Lung and Breast datasets.

Table 4.4 shows the results of the survival prediction models trained on the obtained subspaces $\mathcal{S}$ for different reduction methods.

Type	kSDR	KLR- SDR	PCA	kPCA
Breast BRCA-US	0.58 ± 0.06	$\mathbf{0.60 \pm 0.04}$	0.52 ± 0.04	0.53 ± 0.03
Lung LUAD-US	0.53 ± 0.03	$\mathbf{0.56 \pm 0.03}$	0.55 ± 0.05	0.52 ± 0.02

Table 4.4: Survival prediction results with C-Index score for different dimension reduction methods.

The proposed KLR-SDR outperforms all the benchmark methods with the highest average C-Index after the 20 train-test split. Table 4.4 shows how the standard kSDR is improved by the latent regularization via the KLR-SDR. The problem of survival estimation is more complex due to the large input dimensionality and low sample size (table 2) for this reason it is expected c-index values closer to 0.5 than 1. Nevertheless the KLR-SDR method achieves the highest performance on real data showing its potential for genomic applications.

4.6 Discussion and conclussions

This work extends the concept of Kernel Latent Regularization proposed in the chapter 3 named as *Kernel Latent Regularization for Feature Selection in Tumor Classification* [31] to the Sufficient Dimension Reduction (SDR) method [78]. In this chapter the feature extraction task via Sufficient Dimension Reduction methods is studied instead of feature selection. SDR methods are dimension reduction methods suited for supervised learning task. Particularly regression analysis involves a type of models capable to handle survival analysis problems. For this reason SDR methods are suitable to deal with survival and gene expression data from cancer patients of the International Cancer Genome Consortium [90].
The motivation to use the kernel latent regularization is based on the idea to deal with limited amount of data (small n), as a way to preserve inner-class structure and to account the ignorance about the initial dataset information content. These aspects are observed in the Figure 7 of the toy example where it shows how the standard SDR methods fail to retain the structural distribution of the data after dimension reduction and only capture the discriminant information of the conditional distribution $X \mid Y$. This situation causes firstly the inability to explore sub-clusters within a tumor class after SDR reduction and secondly overfitting since learning a low dimensional representation only from the labels in a context of low sample size may limit the obtained representation to new unseen test samples thus miss generalization capacity.
The standard kernel SDR method uses a target kernel K_y to guide the reduction. In this work it is proposed to replace it with the K_δ kernel as target in order to perform dimensionality reduction from a supervised to an unsupervised approach in the same setting. The toy example visualized in Figure 7 on synthetic data shows how the structure of the data is preserved if the kernel latent regularization is used. In addition, the experiments on synthetic

data for binary classification reveals how the performance of a support vector classifier is improved when the kernel latent regularization is used in the dimension reduction task.
Experiments using real data of gene expression profiles from cancer patients labeled with survival information shows how to learn a low dimensional representation $\mathcal{S}$ regularized by kernel latent regularization which is finally used to train a Cox Regression model for survival prediction. The kernel latent regularization improves the survival prediction in all the datasets compared with the fully supervised case. The proposed Kernel Latent Regularization Sufficient Dimension Reduction method (KLR-SDR) with the mixture parameter δ that peaks the regression performance is compared with kSDR, PCA and kPCA method. The KLR-SDR method outperforms all the benchmark methods showing the potential of the latent regularization module. In this work the kernel-PCA method is used to extract a latent space $\mathcal{Z}$ for the latent regularization term, nevertheless any other unsupervised method can be used such as t-SNE, Autoencoders or PCA.
The obtained results in both synthetic and real data show that learning not only from supervised labels but also from unsupervised latent structure can improve the downstream supervised learning task. Results suggest that latent regularization improves both classification and regression tasks for this reason the proposed kernel latent regularization approach could be generalized to any supervised learning method that uses kernels to guide the learning process.
The Kernel Latent regularization improves a Kernel Sufficient Dimension Reduction method for survival regression problems of cancer patients. The motivation of using internal data structure during training is based on taking account about the ignorance of the initial dataset information content since the low data density may not allow the discovery of inner class subclusters. Moreover, using data structure mixed with labels improves the generalization of a supervised task on new unseen test data.
Kernel methods have shown the potential of improvement even in complex scenarios when the number of features far exceed the number of available samples $d > n$ like cancer genomic data. Further work will consider multi-omic data from patients.

Chapter 5

Unsupervised feature selection with Autoencoders and Kernel Methods

5.1 Introduction

This chapter is devoted to explore and understand feature selection and dimensionality reduction methods from gene expression data of tumor profiles with the objective to discover biomarkers and tumor subtypes via unsupervised machine learning methods. Chapter 03 defines supervised learning problems and feature selection methods to improve a classification task. A main aspect of chapter 03 is the availability of tumor labels y like subtype or survival during the training of the model which allows the application of classifiers like support vector machines. Moreover in Chapter 03 tumor labels are used during the feature selection process thus not only the classification but also the selection is supervised. Nevertheless tumor labels are not always available during the learning process and unsupervised method are needed in this cases [26]. Unsupervised learning is a wide branch of machine learning methods that ranges from dimensionality reduction to clustering. Additionally the high dimensional context of the genomic data from tumor profiles requires to reduce the dimensionality to improve the performance of a downstream unsupervised learning task. One of the questions that this chapter tries to answer is: it is possible to select genes without clinical labels in a way that facilitates the tumor subtype discovery? This chapter formalizes and explains how to select a reduced subset of genes that improves the tumor clustering as a way to guide the discovery of new tumor subtypes [96]. The interpretation of the biological system is done through biomarker discovery with feature selection methods and the subtype discovery through a clustering task.
Each tumor profile is described by more than $d = 10.000$ gene expression features in comparison with a relatively low sample size n which defines a high dimensional input space and further complexity in data analysis. Additionally tumor types presents inner heterogeneity that can be sub-divided in sub-groups with common clinical traits and are named tumor sub-types, thus the tumor subtype discovery is an important task in cancer genomics since it enables clinical doctors to assign a more specific diagnosis and treatments to patients within

the same tumor type [97] [98] [99]. Given the high dimensional space from the input data it is necessary to reduce the dimensionality while preserving the biological interpretation of the system like a reduced gene signature. These reasons motivate this chapter to propose two novel unsupervised feature selection methods for tumor clustering.

As explained in chapter 03 feature selection methods are necessary in cancer genomics since they provide a low dimensional representation of the input data characterized by a selected subset of the input genes providing interpretability of the results and discarding the rest by following an objective function related to improve a learning task [100]. In addition, the selected genes can be used to guide biomarker discovery strategies [24]. The resulting subspace obtained by a feature selection method is described explicitly by a subset of biological features assuming that the initial feature set contains noisy features that can be discarded. A reduced feature subset has benefits in reducing model complexity and in measuring only a reduced set of biomarkers [26].
In an unsupervised problem since the labels y are not available during training it is expected that the selected genes may reveal the data structure which is evaluated in an improvement on clustering of tumor profiles. To guide the feature selection process this work proposes a method that first learn a low dimensional and denoised representation of the input data known as latent space $\mathcal{Z}$ which is learned by an unsupervised neural network known as Autoencoder. Then two strategies are proposed in this chapter. The first is using a Multiple Kernel Learning model to select a subset of gene features with the objective to align as much as possible the resulting distribution from the selected features to a target distribution of the training samples in the learned representation $\mathcal{Z}$. The second strategy, instead of using a Multiple Kernel Learning method for the feature selection task, uses a vanilla autoencoder as a student network to learn a second low dimensional latent space $\mathcal{Z}_v$ in a way that the discrepancy of the data distribution in $\mathcal{Z}_v$ against the one in the latent space $\mathcal{Z}$ is minimized by using the Maximum Mean Discrepancy distance (MMD) [101]. Finally in both proposed methods by doing clustering on the subspace obtained from the selected features it is expected to observe significant clinical attributes associated to each cluster and by this way validate the quality of the selected features.
The main contribution of this chapter are two unsupervised methods able to select genes with clinical relevance from high dimensional gene expression data without the need of having tumor labels.

5.2 Unsupervised Learning

The goal of unsupervised learning is to capture the structure of the data distribution based on similarity between data samples. The key difference between supervised and unsupervised learning is that the latter does not use any response label y during the training of a model. In this line, unsupervised feature selection relies on finding a subset of input features that captures the data structure without using any tumor label.
In the cancer genomics context the gene expression data is assumed to be an independent

random variable represented by a vector x in a space $\mathcal{X} \in \mathbb{R}^d$ where each dimension j of the input vector x_{ij} is a gene and represents the corresponding expression level. The set of n samples S_u is represented by the input matrix $\mathbf{X}$ that group all n input $d-$vectors x_i, $i = 1...n$ where x_i is the ith sample of the dataset. Then given a set S of n samples

$$S = \{x_1, x_2, ..., x_n\} \tag{5.1}$$

the unsupervised learning objective is to describe the structure, the distribution of the data and the similarity between samples in the set S.
The motivation of using unsupervised learning for feature selection on tumor profiles relies on two assumptions. The first assumption states that within a known tumor type class there may exists internal sub-groups and inner heterogeneity of clinical importance. Therefore it is assumed that a tumor dataset may present groups or clusters of tumor samples with high intra similarity that are not described by a clinical label so that the search of these sub groups is desirable. The second assumption is based on the idea that the distribution and structure of the input data has an *intrinsic dimensionality* p lower than the initial input dimensionality d. For this reason it is assumed that it is possible to describe the variablity and structure of the training data in a lower dimension.
In this work the approaches of *clustering* [102], *latent variable models* [60] and *kernel methods* [18] are used to solve the problems of finding subgroups and reduce the dimensionality of the input data. These methods are combined to perform unsupervised feature selection on gene expression data from tumor profiles.

5.3 Clustering

Clustering methods are used to find sub-groups or *clusters* within a set of samples S based on a similarity measurement between samples. The main idea is to partition the initial dataset S in C clusters where each cluster c_i is a subset of samples s_i from the initial sample set. The obtained groups are named *clusters*. The objective is to learn clusters in order to increase the similarity between tumor samples within the same cluster while decrease the similarity between samples from different clusters. This means that after finding the clusters the intra cluster similarity has to be higher than the inter cluster similarity. Clustering methods are useful since tumor types may presents inner heterogeneity thus they can be divided in subgroups. Clustering potentially allows the discovery of new subtypes.
Given a set S of n tumors characterized by d gene expression features the clustering methods define a set C of c subgroups as $C = \{C_1, C_2, ..., C_c\}$ by labeling and assigning each sample x_i to a group C_j. The union of the obtained clusters is equivalent to the full dataset of n samples

$$C_1 \cup C_2 \cup ... \cup C_c = \{1, ..., n\} \tag{5.2}$$

and simultaneously each sample belongs only to one single cluster

$$C_i \cap C_j \quad = \quad \forall\ i \neq j \tag{5.3}$$

which is equivalent to say that clusters are not overlapping [32].

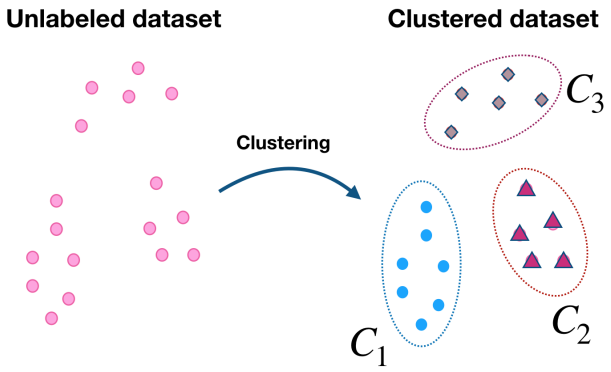

Figure 5.1: Clustering process. Unlabeled samples can be clustered based on similarity.

Figure 5.1 shows a diagram of how samples are clustered by similarity in different groups. In this toy example samples are grouped using the euclidean distance as a measure of similarity so similar samples are grouped together in the same cluster.

5.3.1 k-means method

The k-means algorithm [103] is one of the most used clustering methods in unsupervised learning. This algorithm has been used in different domains such as computer vision [104] and Graph data structured as networks [105]. Moreover, k-means has been applied in cancer genomic application such as survival analysis [106]. It can be used in cases where the features are defined in the domain of continuous real-valued numbers. As defined previously, clustering methods use a similarity measurement between samples. For instance, the k-means method can use as similarity measurement, the Euclidean Distance for any pair of samples $x_i, x_{i'}$ as

$$d\left(x_i, x_{i'}\right) = \sqrt{\sum_{j=1}^{d} \left(x_{ij} - x_{i'j}\right)^2} = \|x_i - x_{i'}\| \tag{5.4}$$

Note that the shorter the distance, the higher the similarity. Since the goal of clustering is to group together in the same cluster samples with high similarity then the objective function $\mathcal{L}$ to minimize is defined as

$$\begin{aligned} \underset{min}{\mathcal{L}(C)} &= \frac{1}{2} \sum_{j=1}^{c} \sum_{i \in C_j} \sum_{i' \in C_j} d\left(x_i, x_{i'}\right) \\ &= \frac{1}{2} \sum_{j=1}^{c} \sum_{i \in C_j} \|x_i - \bar{x}_j\| \end{aligned} \tag{5.5}$$

where x_i and x_i' are every pair of samples belonging to the same cluster C_j, $\bar{x}_j$ is the centroid or mean vector associated to cluster C_j. The objective function $\mathcal{L}(C)$ is minimized

by assigning n samples among the c clusters in such a way that the mean dissimilarity of samples within a cluster C_j with respect to the cluster centroid is minimized. By optimizing the objective function in equation 5.5 the k-means method aims to minimize the total inner variance of the resulting clusters with respect to their corresponding centroids. The solution of the optimization problem results in a local minima by the Elkan's algorithm [107].

5.3.2 Evaluation metrics in Clustering

Since the objective of this chapter is to perform unsupervised feature selection, the clustering task is used as a downstream task which means that it is done after the selection step in order to evaluate the quality of the selected features. K-means clustering method [103] is applied on the tumor samples characterized by p selected gene features for which the true labels are known. The quality of clusters is evaluated by the Adjusted Rand Index [108] and is done by comparing the learned cluster label c_i with the ground truth y_i labels related to the tumor subtype and provided a posteriori as clinical data only for evaluation purposes. It is important to remark that the ground truth labels are never used to select the features neither learning the cluster and are only used to measure how well the k-means groups the tumor samples. It is computed as

$$\text{Rand Index} = \frac{\text{A} + \text{B}}{\text{A} + \text{B} + \text{C} + \text{D}} \tag{5.6}$$

where A is the number of tumor sample-pairs assigned to the same cluster and belonging simultaneously to the same tumor subtype, B is the number of tumor sample pairs assigned to different clusters and simultaneously belonging to different tumor subtypes, C is the number of tumor sample pairs assigned to the same cluster but belonging to different tumor subtypes and D is the number of tumor sample-pairs assigned to different clusters and belonging to the same tumor subtype. The Rand Index can be thought as a clustering accuracy and takes real values from 0 to 1 where a value close to 0 means a random and non informative clustering results regarding to the ground truth clinical labels. When the Rand Index is close to 1 it means that almost every cluster is populated with tumor samples of the same subtype which is a desired score.

5.3.3 Clustering example

Let the Lung Cancer dataset from the ICGC data repository [16] used in the previous chapter with $n = 135$ samples. The dataset is composed of two classes each one associated to a lung cancer subtype: Squamus Cell and Adenocarcinoma. Therefore samples can be labeled as $y \in \{-1, 1\}$ where this labels are defined as ground truth nevertheless these are not used during the unsupervised learining process and are available only for further validation of the clustering result. The number of genes used in this example are just two in order to facilitate the visualization for this reason $d = 2$. The genes are the *RPLP1* and *GAPDH* and are the ones with the highest variance genes in all the lung cancer dataset. The data matrix $\mathbf{X}_{(n,d)}$ is $n = 135$ and $d = 2$.

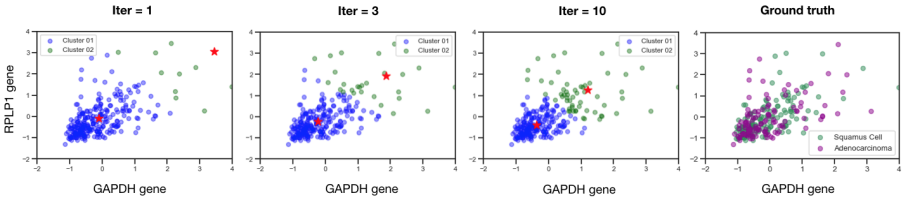

Figure 5.2: Clustering process.

Figure 5.2 shows a scatter plot of the data distribution along the two gene expression features where each dot is a sample. From left to right the first three plots show the resulting clusters and their corresponding centroids obtained from a K-means algorithm for different iteration number. It is observed how the centroids and the cluster labels evolve across each iteration of the k-means algorithm in a way to minimize the cluster inner variance. The third plot from the left shows the final result of the k-mean method after 10 iterations. The plot on the right shows the ground truth labels. It can be seen visually that the assigned cluster labels by the k-means do not overlap enough with the ground truth labels. To measure the quality of the clusters the Rand index is used with different number of clusters c.

c = 2	c = 3	c = 4	c = 5
0.11	0.17	0.18	0.17

Table 5.1: Rand index for different number of clusters.

Table 1 details the quality metrics used to evaluate the clustering results. The Rand Index in figure 5.2 is 0.11 with two clusters $c = 2$ and increases after $c = 3$. Ideally the best clustering results is the one with the Rand Index close to 1.

5.4 Dimensionality reduction with Autoencoders

Dimensionality reduction is a branch of machine learning models designed to learn a low dimensional subspace $\mathcal{Z}$ known as *latent space* from the input space $\mathcal{X}$ via a mapping function $f(\mathcal{X})$ as detailed in figure 5.3. The key idea is to extract features from a linear or non-linear combination of the input ones. The objective of learning new features relies in finding a denoised and low dimensional subspace $\mathcal{Z}$ that captures the salient features of $\mathcal{X}$ in a lower dimension. Learning a low dimensional latent space $\mathcal{Z}$ from gene expression tumor profiles may improve downstream machine learning tasks like clustering when compared with the high dimensional input space $\mathcal{X}$.

The mapping function $f()$ can be learned by linear and non-linear methods. One of the most common linear methods for dimensionality reduction is th Principal Component Analysis (PCA) which aims to project the input space $\mathcal{X}$ into a subspace $\mathcal{Z}$ that best captures

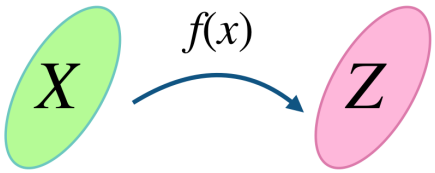

Figure 5.3: Dimensionality reduction: input data X is mapped to a low dimensional latent space $\mathcal{Z}$ via a function $f(X)$.

the variance of the input data. Another well known non-linear method is the kernel-PCA, presented in the previous chapter as part of the KLR-FS method [31]. Finally the Autoencoder method, a neural-network based approach, is a non-linear dimensionality reduction method widely used in signal processing, computer vision and natural signal processing. One of the main advantages of the autoencoder method is its ability to learn complex mapping functions $f(x)$ by learning the neural network parameters.

5.4.1 Related Work

Reducing the dimensionality of gene expression data can be achieved by feature selection [26] and feature extraction methods [109].
Feature extraction is the construction of a reduced subset of l new features obtained from a linear or nonlinear combination of the initial set of features. Neural Networks have gained popularity for feature extraction and dimensionality reduction lead by the Autoencoder model which is based on nonlinear transformations [75]. The reduced dimensional space $\mathcal{Z}$ and the extracted features are known as *latent space and latent features* respectively [2]. Autoencoders have been used in biomedical problems to integrate multi-omic data like gene expression, methylation and microRNA to predict Liver cancer prognosis [106]. In a similar way autoencoders have been used to learn meaningful representations from gene expression data of Breast cancer patients and then identify tumor subtypes [110]. In addition, Autoencoders have been applied to learn a latent space from somatic mutation data of the pan-cancer landscape showing improvements in the clustering performance [111] [67]. Moreover, Variational Autoencoders (VAE) have been trained on DNA Methylation data from Lung Cancer patients [112] and on gene expression data of pan-cancer tumor samples to learn meaningful representations for supervised and unsupervised tasks [111] [113]. These works attest of the important role and capacity of Autoencoders for feature extraction and dimensionality reduction on molecular data from tumors.

5.4.2 Autoencoders

Autoencoders (AEs) are the dimensionality reduction method used in this chapter. AEs are feed forward artificial neural networks (ANNs) and have the objective to learn two functions,

an encoder $f(x) = z$ and a decoder $q(z) = x$ as detailed in figure 5.4. The encoder is a non-linear function that maps the input domain $\mathcal{X}$ of dimension d to a latent space $\mathcal{Z}$ of lower dimension l. On the other side, the decoder is a function designed to reconstruct the samples from the latent space $\mathcal{Z}$ to the input space $\mathcal{X}$.

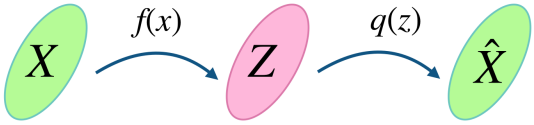

Figure 5.4: Autoencoder data transformations.

The encoder is forced to learn a function that captures the salient features from $\mathcal{X}$ and maps to $\mathcal{Z}$ [2]. The encoder function is defined as $\boldsymbol{z} = f(\boldsymbol{x})$ and the decoder as $\tilde{\boldsymbol{x}} = q(\boldsymbol{z})$. The samples at the latent space are expressed by $\boldsymbol{z}$ while $\tilde{\boldsymbol{x}}$ represents the reconstructed samples by the decoder function lying on $\mathcal{X}$. During training the autoencoder has to minimize the expectation of the following loss function

$$L(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \frac{1}{n} L(\boldsymbol{x}, q(f(\boldsymbol{x}))) \tag{5.7}$$

where L penalizes $q(f(\boldsymbol{x}))$ when it is different from $\boldsymbol{x}$. The loss function can be computed by the Mean Squared Error (MSE) expressed as

$$L_{MSE}(X, \hat{X}) = \sum_{i=1}^{n} ||x_i - \hat{x}_i||^2$$

The encoder f and decoder q functions are expressed as [114]

$$\begin{aligned} \boldsymbol{z} = f(\boldsymbol{x}, \mathbf{W}_f, \mathbf{b}_f) = &\quad \sigma(\mathbf{W}_f \boldsymbol{x} + \mathbf{b}_f) \\ \hat{\boldsymbol{x}} = q(\mathbf{z}, \mathbf{W}_q, \mathbf{b}_q) = &\quad \sigma(\mathbf{W}_q \mathbf{z} + \mathbf{b}_q) \end{aligned}$$

The encoding function is $f(\cdot, \mathbf{W}_f)$ and the decoding $q(\cdot, \mathbf{W}_q)$. The expression $\sigma(\cdot)$ is the activation function of the neurons of the ANN. The matrices $\mathbf{W}$ and $\mathbf{b}$ are the network parameters to learn with the objective to minimize the estimated loss function and represent the weights and biases of the encoder and decoder functions respectively. The autoencoder weights are trained by using the back-propagation algorithm. The optimizer used to learn the parameters of the network is the Adaptive Moment Estimation (Adam) [115] which is based on the short term gradient mean with adaptive learning rate to speed up the learning process.

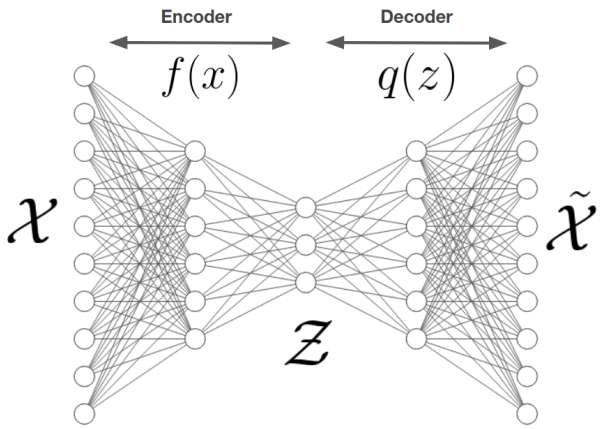

Figure 5.5: Architecture of an autoencoder.

In order to force the encoder to learn a useful representation on the latent space and to avoid the AE to just copy $\hat{\boldsymbol{x}}$ from $\boldsymbol{x}$ different regularization strategies are implemented. First, a regularization term using the L_2 norm is imposed on the weights Wf_{ij} and Wq_{ij} and added to the loss function L. The regularization hyper-parameter is $\beta \geq 0$ as follows

$$L_r = L\left(\boldsymbol{x}, q\left(f\left(\boldsymbol{x}\right)\right)\right) + \beta \sum_i \|w_i\|^2 \tag{5.8}$$

The regularization term avoid both f and q to have large weights and leads to learn a simpler model, in consequence this reduces the overfitting of the trained model. A second regularization strategy that helps to improve even more the generalization capacity of the model is Batch Normalization (BN) [116] which consists to perform normalization at each mini-batch iteration during training. During Autoencoder training, the training data is splited in multiple subsets named mini-batches where the autoencoder parameters w are adapted on each mini-batch instead of learning from the full dataset. In order to minimize the Loss function its gradient is computed on each mini-batch and used to tune at each iteration the network parameters. The training data passes through the autoencoder multiple times where each time is known as *epoch*. Training is done after multiple epochs when the loss function converges to a desired local optima.
In this work autoencoders are used to learn meaningful and low dimensional representations from gene expression data from tumor profiles.

5.4.3 Example

Let the dataset $\mathbf{X}$ of Lung Cancer from the ICGC data portal composed of $n = 906$ samples and $d = 17233$ gene expression features. The samples of the Lung Cancer dataset are composed by two ground truth tumor subtype labels: Squamus Cell subtype (labeled as LUSC-US) and Adenocarcinoma subtype (labeled as LUAD-US). In this thesis two types of

Autoencoders with two different architectures are used plus the Kernel-PCA. To analyze the performance of the dimensionality reduction method a downstream unsupervised learning task is applied on the learned latent space. The downstream task in this example is clustering and the latent space obtained by each method is analyzed based on the capacity to cluster tumor samples accordingly to clinical data.

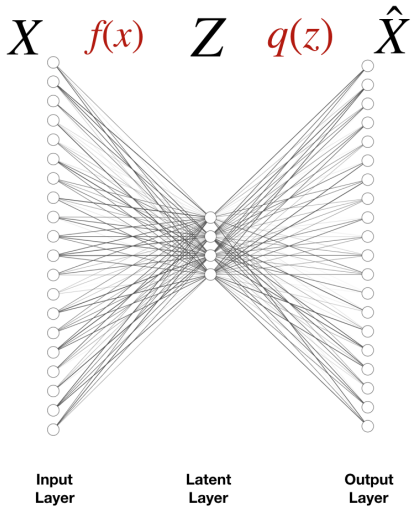

Figure 5.6: Vanilla Autoencoder architecture.

One of the autoencoder architecture used is the Vanilla Autoencoder. This model is composed by just one hidden layer located in the bottleneck of the autoencoder. This corresponds to an encoder and decoder without any hidden layer. The Vanilla Autoencoder relies on its simplicity and is a reference for a low complexity neural network. Nevertheless when implementing this model no regularization term were introduced on its weights, therefore there is not control over the parameter domain which may cause the model to overfit.
Another possible autoencoder architecture is a deep regularized one. In this example the model has two hidden layers in both the encoder and decoder functions and a penalty on the encoder and decoder weights. The regularization shrinks close to zero the weights that do not contribute significantly to perform the encoder and decoder tasks.

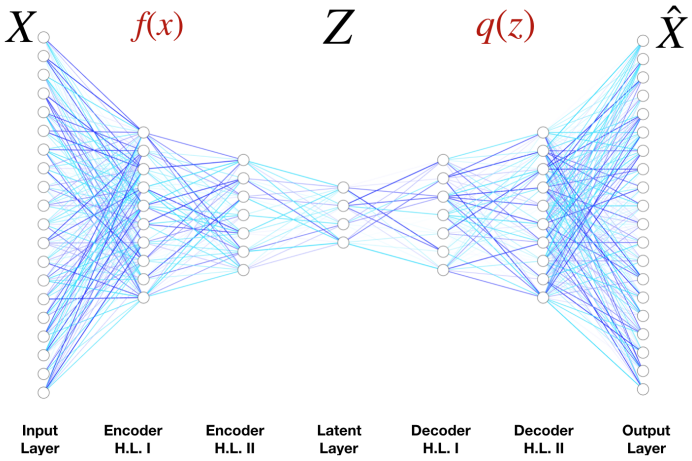

Figure 5.7: Deep Regularized Autoencoder architecture.

The architecture is named deep since it has multiple hidden layers in the encoder and decoder networks. By stacking multiple hidden layers the features extracted in the latent space might be more complex and compact when compared with the vanilla autoencoder. In addition, multiple hidden layers in the decoder may help the reconstruction process from the latent space to the input space.
Once the three models (kPCA+2 autoencoders) are trained their corresponding latent spaces can be compared. By applying a PCA on each latent space and keeping just the first two components for visualization purposes the sample distribution of each class can be observed. The Vanilla AE and the kernel-PCA tend to group together the samples from both classes in a single cluster. On the other side the Deep Regularized AE captures the distribution of the Raw data and group the samples of each tumor subtype in two different regions. Visually it can be observed how the Deep Regularized AE outperform the other two methods by how the samples of the same class seems to be grouped.

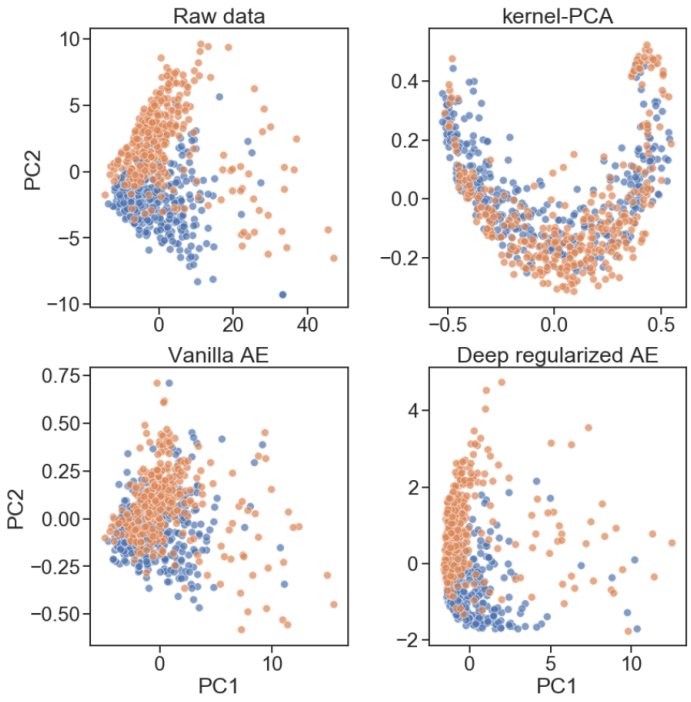

Figure 5.8: Visualization of the latent space of each dimensionality reduction method. Blue dots: Squamus Cell subtype. Orange dots: Adenocarcinoma subtype

Finally to evaluate quantitatively the quality of each dimensionality reduction method a k-means clustering algorithm is applied on each obtained latent space as a downstream task. With the resulting clusters the Rand Index is computed by comparing the obtained cluster labels and the ground truth ones corresponding to the tumor subtypes.

Method	Features	$c = 2$	$c = 3$	$c = 4$	$c = 5$
Raw	17233	0.001	0.017	0.144	0.166
kPCA	50	**0.034**	0.046	0.031	0.031
Vanilla AE	50	0.001	0.016	0.032	0.029
Deep Reg. AE	50	0.010	**0.561**	**0.328**	**0.321**

Table 5.2: Rand index for different number of clusters.

Table 3.2 shows the resulting Rand Index for different values of clusters $c = 2, 3, 4, 5$. It is observed that the maximum rand index is obtained by the Deep Regularized Autoencoder for $c \geq 3$ that outperformed significantly the other methods. This evidence how the deep and regularized architecture of the autoencoder improves the structure of the learned latent space in a way that the resulting clusters correlates significantly to the tumor subtypes. Moreover the rand index results of the deep regularized autoencoder are higher than the ones obtained

with the raw data. These results suggest that the best method to extract structure from the high dimensional input data is the Deep Regularized Autoencoder.

5.5 Unsupervised Feature Selection

Dimensionality reduction via latent variable models such as Autoencoders or kernel-PCA reduce the dimensionality by creating a new reduced subset of features from the linear or non linear combination of the input ones. The latent features obtained may capture the inner structure of the input data and the salient characteristics about its distribution. Nevertheless dimensionality reduction via latent variable models implies that the obtained features are not interpretable since these are built from the combination of all the input features. Therefore to reduce the dimensionality of the input data by maintaining the interpretability feature selection methods are used.
Feature selection as explained in the previous chapter consists in the selection of a subset $P = \{1, 2, 3, 4, ..., p\}$ of p features from the input feature set $D = \{1, 2, 3, 4, ..., d\}$ where $p < d$. As explained in chapter 03, when sample labels y are available supervised feature selection is used with the objective to improve a supervised learning task. Nevertheless big quantities of unlabeled tumor data have been generated in biology as explained in the clustering section. Each tumor type may present inner-heterogeneity suggesting that a tumor type or subtype can be de-composed in sub-clusters. It is desired to preserve the inner structure of the training data and to be able to explore subtypes after the feature selection. Additionally the input dimensionality of transcriptome data is $d > 10.000$ with d indicating the number of gene expression features therefore reducing the input dimensionality is needed.
This section first introduces two unsupervised feature selection techniques and then proposes two novel approaches based on kernel methods and autoencoders.

5.5.1 Related Work

As detailed in chapter 3 the Multiple Kernel Learning (MKL) have been used for gene selection in supervised problems with the objective to improve the classification between tumor types [117][118]. Feature selection has been applied also on multi-omic data like Gene Expression, Methylation and miRNA using the Minimum redundancy - maximum relevance (mRMR) [119] method to predict survival rate on Glioblastoma Multiforme patients [49]. Another study proposes a stable feature selection method for high-dimensional RNA-seq data while applying an ensemble L_1-norm support vector machines to reduce irrelevant features [120] and classify tumor stages of renal clear cell carcinoma. In addition, feature selection by Elastic Net [121] has been proposed to select genes linked to the Triple Negative Breast Cancer subtype [122]. The papers described above show the potential and necessity of supervised feature selection methods for gene selection on cancer molecular data.
Nevertheless, labeled data is not always available and the selection of genes is needed for unsupervised tasks such as clustering since tumor types may present heterogeneity and each cluster can present different clinical properties. The problem of unsupervised gene selection

is faced by Unsupervised Feature Selection methods for Clustering [123]. Multi-Cluster Feature Selection (MCFS) [124] is an unsupervised model proposed to select the features that preserves the cluster structure of the original data and has been applied on micro-RNA data [125]. Also the Sparse k-Means (SKM) method [126] has been proposed to weight each feature based on the partition of data and by this way a subset of features is selected by penalizing weights with the L1 norm. Moreover, an unsupervised spectral method (SPEC) [127] has been proposed to determine relevant genes on acute lymphoblastic leukemia [128]. Feature selection and feature extraction methods both have shown potential to reduce the dimensionality of tumor data. In this chapter we propose two methods that combines both strategies by learning a low dimensional latent space via feature extraction from the input data and then selecting the gene expression features that approach the most to the resulting learned latent representation. By doing so it is expected the selected features to retain the same information as the latent variables. The proposed methods presented in this work are compared with the SKM and SPEC detailed below as benchmark.

5.5.2 Sparse K-means method

The first benchmark method is the Sparse K-Means (SKM) [126]. The SKM computes via an optimization problem feature weights $\mathbf{w} = [w_1, ..., w_d]$ and applies a lasso-type L1 penalty $||w||_1 < \alpha$ to select the most important features while doing k-means clustering. The feature weights are a measurement of the variable importance in clustering. The SKM method is based on the K-means type family of algorithms and assigns a larger weight to the features that have a smaller sum of intra-cluster distances and a smaller or zero weight to the features with a high intra-cluster distance.

5.5.3 SPEC method

The other benchmark method is the Spectral Feature Selection for unsupervised learning (SPEC) which is based on spectral graph theory. SPEC uses a pairwise similarity matrix $\mathbb{S}$ between samples to build a graph $\mathbb{G}$ where each node is a sample and each edge is the similarity measurement. The idea with SPEC is to select the features that are consistent with the graph structure. The objective of SPEC is to select features that gives similar values to samples that are near each other in the graph. A graph $\mathbb{G}$ can be built from the pairwise similarity obtained from $\mathbf{X}$. Then the SPEC method makes a feature ranking based on the Normalized Cut of the graph $\mathbb{G}$ by using the corresponding Laplacian matrix from the graph.

5.6 Proposed unsupervised feature selection methods

To perform unsupervised feature selection two novel methods are proposed in this chapter. The proposed methods are Latent Kernel Feature Selection (LKFS) method and Latent Maximum Mean Discrepancy Feature Selection (LMMD-FS).

5.6.1 Proposed method I: Latent Kernel Feature Selection

Given a set of n tumor samples characterized by d gene expression features the data is contained in a $\mathbf{X}_{nd}$ expression matrix. It is desired to select a subset of p features from d. To achieve this goal first an autoencoder model is trained and a latent space $\mathcal{Z}$ of dimension l is obtained where $l << d$. This latent space is assumed to capture the key information with less noise as the initial representation space. Then using the set of samples n projected in the latent space a gaussian kernel is used to build a kernel matrix K_z as target kernel. The target kernel matrix K_z measures similarity between pair of samples in the latent space.

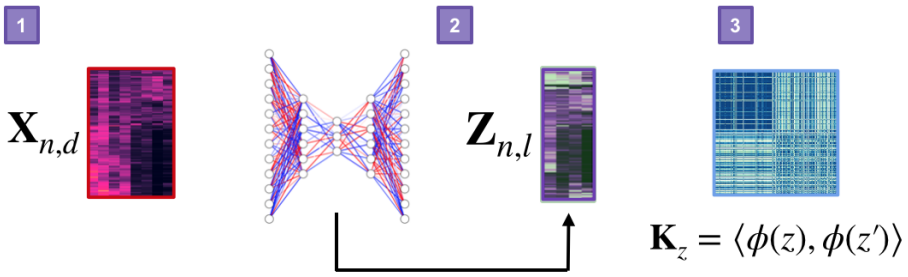

Figure 5.9: Pipeline to build the unsupervised target kernel. First an autoencoder is used to reduce the dimensionality of the data. Second a latent space is obtained after training. Finally a kernel matrix K_z is built by using the projected samples on the latent space.

To learn the latent space an architecture for both the Encoder and the Decoder functions is proposed. Starting from the Encoder the input layer of dimension d is fully connected to a hidden layer HL_1 of 200 neurons. Then HL_1 is fully connected to a second hidden layer HL_2 with 100 neurons. Finally HL_2 is connected to the latent hidden layer HL_z. The latent representation Z has a lower dimension $l = 50$ in comparison to the input dimension d and describes with less noise the original data. Symmetrically and starting from the latent layer HL_z the Decoder Function has one hidden layer HL_3 of 100 neurons followed by one HL_4 of 200 neurons and finally the output layer of dimension d. The training of the autoencoder has been determined by the following hyperparameters: ReLu hidden neurons, linear neurons in the lat decoder layer, $L_2 = 0.0005$ norm on the encoder model, learning rate $= 0.0001$, epochs $= 50$, bach size $= 32$, and validation set $= 20\%$.
Using the sample set $\{z_1, ...z_i, ...z_n\}$ in Z a similarity matrix K_z is defined based on a gaussian kernel and it is used as target matrix K_T kernel by the following Multiple Kernel Learning method as detailed in figure 5.10. After defining the $K_T = K_z$ matrix Multiple Kernel Learning method is used to select features by having as reference the target matrix K_z from the representation obtained with the autoencoder.

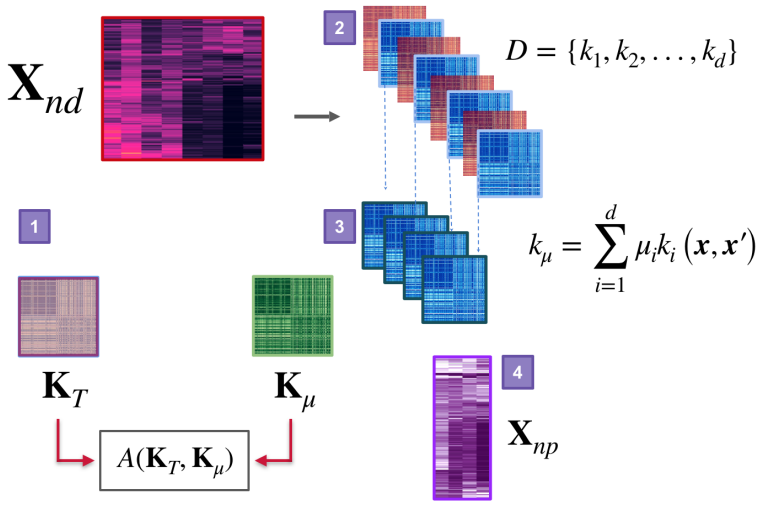

Figure 5.10: Feature Selection with Multiple Kernel Learning. (1) First define a target kernel K_T. (2) Build d feature-wise kernels. (3) Perform a linear combination of all the kernels by maximizing the aligment of the resulting kernel with the target one K_T. (4) The kernels used in the final solution are the ones that defines which features select.

From $\mathbf{X}_{nd}$ a set of d feature-wise kernels is built producing one kernel per feature. Then by using the MKL model described in the previous chapter a reduced subset of p kernels is iteratively selected and combined to build a K_μ kernel matrix that increases the alignment $A(K_\mu, K_T)$ as detailed in figure 5.11. The kernel k_μ is built from a linear combination of single kernels k_i [43] and is expressed as

$$\mathbf{k}_{\boldsymbol{\mu}}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=1}^{n} \mu_i \mathbf{k}_i(\boldsymbol{x}, \boldsymbol{x}'), \mu_i \geq 0 \tag{5.9}$$

The vector μ represents the weight μ_i of each kernel k_i and reflects the relative importance of each kernel in the final solution $\mathbf{k}_\mu$. There are many possible objective functions to optimize while learning the MKL task, nevertheless in this work the MKL model is built with the objective to maximize the alignment of the resulting kernel matrix $\mathbf{K}_\mu$ with the target kernel matrix $\mathbf{K}_T$. Particularly, the target kernel proposed in this method is $\mathbf{K}_z$ built from the samples z_i lying in the latent space $\mathcal{Z}$ defined by the Autoencoder thus the resulting alignment to optimize is $A(\mathbf{K}_\mu, \mathbf{K}_z)$. Starting from a set of feature-wise kernels $D = [K_1, K_2, ..., K_d]$ the greedy MKL algorithm chooses at the first iteration the kernel K_i with highest $A(K_i, K_z)$. Then it will start adding at each iteration a new kernel to the solution that improves as much as possible the current alignment until $A(K_i, K_z)$ stop increasing. The final vector μ is sparse with only $\mu_i > 0$ for every K_i selected during the MKL process.

Only the feature-wise kernel matrices that increases the alignment $A(K_\mu, K_z)$ are included in the final kernel K_μ. This approach leads to a sparse solution where the non-zero values of the μ vector indicates the feature importance on the result. Features are selected by an unsupervised strategy that best align the representation learned from the autoencoder. This

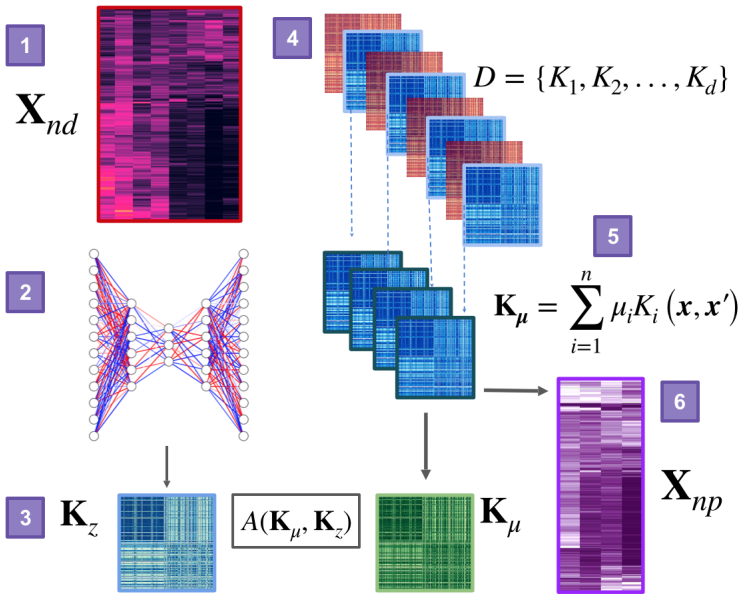

Figure 5.11: Pipeline of the proposed method. First starting from the raw data (1) an autoencoder is trained (2) and a latent space learned. Then a K_z kernel matrix built (3) using the sample set projected on the latent space. Finally feature-wise kernels matrices are built (4) and combined by MKL (5) to obtain a K_μ kernel matrix by improving the aligment $A(K_\mu, K_z)$. The result is a X_{np} matrix characterized by a subset of p features associated to the feature-wise kernels selected by MKL as $\mu > 0$.

method has been named Latent Kernel Feature Selection (LKFS) [129]. Figure 5.11 shows a diagram of the proposed method.

5.6.2 Proposed method II: Latent Maximum Mean Discrepancy Feature Selection

The LKFS method presented previously uses the latent space of the autoencoder to guide the training of the MKL mode, therefore the selected features by MKL will describe a structure similar to the one obtained from the latent space of the autoencoder. This lead to the concept of a method with two models where one model learns a latent space and the other model follows the structure of the first one. This approach is known as *Teacher-Student* [130] where the deep autoencoder is considered the teacher method. In this section another novel unsupervised feature selection method is proposed. This new method is based on the teacher-student approach where the deep autoencoder is still being used as teacher method and a vanilla autoencoder is used as the student one instead of the MKL algorithm. The motivation to use a Vanilla Autoencoder as student instead of a MKL model relies in the fact that training a MKL task involves the combination of thousands of kernels while a Vanilla AE just involves one single overparametrized model, therefore the vanilla autoencoder makes

the whole approach simpler. In addition, instead of using the Kernel Target Alignment to measure the discrepancy between the teacher and student representations the Maximum Mean Discrepancy measurement is used [101]. The selection task is based on a ranking of the input features computed via the L_1 norm of the weights in the first and unique layer of the encoder in the vanilla autoencoder. The reason why the vanilla autoencoder is not used alone to just learn the latent space and rank the input features simultaneously is because it has not enough capacity to learn a meaningful latent space with a single layer in comparison with the deep regularized one. Additionally keeping just a single layer for the encoder section makes easier to study the influence of each input feature in the resulting latent space. For this reason in the proposed method the deep autoencoder is used to learn the target low dimensional latent space and the vanilla autoencoder for the feature selection task.
The proposed method is named *Latent Maximum Mean Discrepancy Feature Selection* (LMMD-FS) and aims to select features by an unsupervised approach. As explained previously, the deep autoencoder is a method capable to learn a meaningful latent space from gene expression profiles. To learn the latent space $\mathcal{Z}$ the encoder function $f(x)$ is used where $f(x) = z$ and the training is performed with the objective to minimize the reconstruction error $L_{\text{MSE}}(x, \tilde{x})$ of the autoencoder as $f(q(z)) = \tilde{x}$. Simultaneously the student model composed by a Vanilla Autoencoder is trained with the same input data, the same reconstruction loss function and additionaly the objective to minimize the Maximum Mean Discrepancy between the latent space $\mathcal{Z}$ of the Deep Regularized Autoencoder and the one $\mathcal{V}$ associated to the Vanilla Autoencoder.
The Maximum Mean Discrepancy (MMD) is a measure of distance between distributions projected in a Reproducing Kernel Hilbert Space (RKHS) and belongs to the family of models known as *Kernel Mean Embeddings* [28]. The MMD metric has been used to improve the training of several machine learning models from Generative Adversarial Networks [131] and Autoencoders [132], therefore is a useful tool for representation learning [2]. The idea of Kernel Mean Embeddings is to extend the feature maps $\phi(x)$ presented in the kernel methods section of chapter 03 to probability distributions by representing each distribution $\mathcal{Z}$ as a mean distribution

$$\phi(\mathcal{F}) = \mu_{\mathcal{F}} := \int_{\mathcal{Z}} k(\mathbf{z}, \cdot) \mathrm{d}\mathbb{P}(\mathbf{z}) \tag{5.10}$$

where $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ is a positive definite and symmetric kernel function. Equation 3.10 transforms a distribution $\mathcal{Z}$ to an element in the hilbert space $\mathcal{H}$ (a RKHS associated to the kernel function k [28]) which is the kernel mean embedding $\mu_{\mathcal{F}}$ of distribution $\mathcal{Z}$. Then given two sample sets Z and V sampled from two distributions $\mathcal{Z}$ and $\mathcal{V}$ the distance between them is equivalent of the distance of their corresponding mean embeddings in the RKHS.

$$\text{MMD}[\mathcal{H}, Z, V] = \|\mu_{\mathcal{Z}} - \mu_{\mathcal{V}}\|_{\mathcal{H}} \tag{5.11}$$

For this reason if $\mathbb{Z} = \mathbb{V}$ then $\|\mu_{\mathcal{Z}} - \mu_{\mathcal{V}}\|_{\mathcal{H}} = 0$. This means that if the two distributions are close in the input space their corresponding mean embeddings are close in the Hilbert space as well.

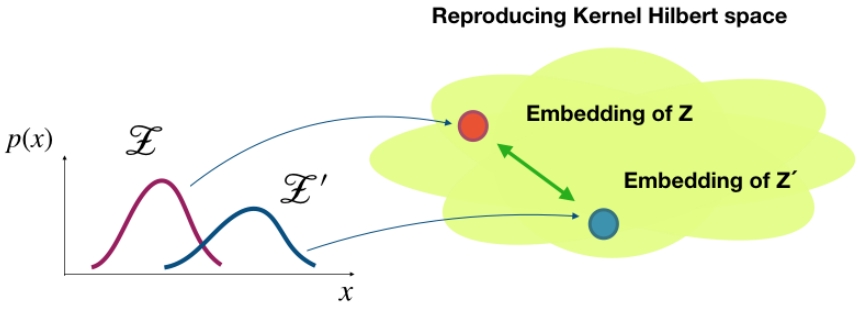

Figure 5.12: Distance between Kernel Mean Embedding of distributions in a Reproducing Kernel Hilbert Space.

On the other side, if the two distributions are distant between each other then $\mathbb{Z} \neq \mathbb{V}$ and $\|\mu_{\mathcal{Z}} - \mu_{\mathcal{V}}\|_{\mathcal{H}} > 0$. One of the key advantages of using this approach relies in the fact that estimating the kernel mean embedding is easier than the estimation of the distribution itself [101]. To compute the distance between distributions the *Maximum Mean Discrepancy* (MMD) metric is used. This metric computes the distance between the corresponding mean embeddings of distributions in the Hilbert space $\mathcal{H}$. To compute the squared MMD metric in terms of a kernel function k between the distributions $\mathcal{Z}$ and $\mathcal{V}$ in the Hilbert space $\mathcal{H}$ then the following expression has to be estimated

$$\begin{aligned} \mathrm{MMD}^2[\mathcal{H}, Z, V] &= \|\mu_z - \mu_v\|^2_{\mathcal{H}} \\ &= \langle \mu_z, \mu_z \rangle_{\mathcal{H}} + \langle \mu_v, \mu_v \rangle_{\mathcal{H}} - 2 \langle \mu_z, \mu_v \rangle_{\mathcal{H}} \\ &= \mathbf{E}_{z,z'} \langle \phi(z), \phi(z') \rangle_{\mathcal{H}} + \mathbf{E}_{v,v'} \langle \phi(v), \phi(v') \rangle_{\mathcal{H}} - 2\mathbf{E}_{v,y} \langle \phi(v), \phi(v) \rangle_{\mathcal{H}} \end{aligned} \tag{5.12}$$

where $\langle \phi(z), \phi(z') \rangle_{\mathcal{H}} = k(z, z')$ and $\langle \phi(v), \phi(v') \rangle_{\mathcal{H}} = k(v, v')$ [101]. The expression $\mathrm{MMD}^2[\mathcal{H}, Z, V]$ is equivalent to the following

$$\mathrm{MMD}^2[\mathcal{H}, \mathbb{Z}, \mathbb{V}] = \mathbb{E}_{z,\tilde{z}}[k(z, \tilde{z})] - 2\mathbb{E}_{z,v}[k(z, v)] + \mathbb{E}_{v\tilde{v}}[k(v, \tilde{v})] \tag{5.13}$$

where $z, \tilde{z} \sim \mathcal{Z}$ and $v, \tilde{v} \sim \mathcal{V}$ are samples belonging to distributions in the latent space of the deep regularized and vanilla autoencoder respectively. Given the i.i.d. sample vectors $\mathbf{Z} = \{z_1, z_2, \ldots, z_n\}$ and $\mathbf{V} = \{v_1, v_2, \ldots, v_m\}$ with n and m samples respectively obtained from $\mathcal{Z}$ and $\mathcal{V}$, the empirical estimation of the squared MMD score is computed as

$$\begin{aligned} \widehat{\mathrm{MMD}^2}[\mathcal{H}, Z, V] = &\frac{1}{m(n-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(z_i, z_j) + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k(v_i, v_j) \\ &- \frac{2}{mn} \sum_{i=1}^{n} \sum_{j=1}^{n} k(z_i, v_j) \end{aligned} \tag{5.14}$$

In this work $m = n$. To take advantage of the MMD metric for feature selection the proposed method has two models. First a deep regularized autoencoder that learns a well defined latent space $\mathcal{Z}$ by minizing the reconstruction loss $\text{MSE}_{\text{DRA}}(x, x')$. Simultaneously a vanilla autoencoder is trained with the objective of minimizing a multi-loss objective function composed by two terms. The first loss term of the vanilla autoencoder is the reconstruction loss as $\text{MSE}_{\text{VAN}}(x, x')$ between the input and reconstructed vectors. The second loss component of the vanilla autoencoder is the distance betwen the learned latent spaces distributions $\mathcal{Z}$ and $\mathcal{V}$ of both autoencoders as $\text{MMD}^2[\mathcal{H}, Z, V]$. Since the vanilla autoencoder has just one single layer the reconstruction loss is easily back-propagated to the weights of the first layer which is connected to the input features. For this reason during training the L1 norm is imposed to penalize large unnecessary weights on the first layer of the Vanilla AE. Then the objective is to minimize the deep regularized autoencoder loss defined as

$$\mathcal{L}_{DRA} = \text{MSE}_{DRA}(\mathcal{X}, \mathcal{X}') \tag{5.15}$$

where MSE_{DRA} is the reconstruction loss, and to minimize the total loss of the vanilla autoencoder defined as

$$\mathcal{L}_{VAN} = \text{MSE}_{VAN}(\mathcal{X}, \mathcal{X}') + \text{MMD}^2[\mathcal{H}, Z, V] + \lambda \sum |w_i| \tag{5.16}$$

where MSE_{VAN} is the reconstruction loss of the vanilla autoencoder, the term $\sum |w_i|$ the $L1$ penalization on the weights, λ a parameter to tune the strength of the selection and $\text{MMD}^2[\mathcal{H}, Z, V]$ is the MMD squared distance between the distributions obtained from both latent spaces. Finally the total loss to be minimized is expressed as

$$\mathcal{L}_{\text{LMMDFS}} = \mathcal{L}_{DRA} + \mathcal{L}_{VAN} \tag{5.17}$$

since both autoencoders are trained together simultaneously. Once both autoencoders are trained the feature selection task is performed by defining a ranking on the input features. The ranking is based on the norm of the weights in the first layer of the Vanilla Encoder $f(x)$ by computing the $L1$ norm of the first-layer weight-matrix $r = \|W\|_1 = \sum_i^l |w_i|$ [130]. With the feature ranking the top p features are selected from r. By this mean p features are selected from the d input ones with the combination of a Deep Regularized AE and a Vanilla AE coupled by the MMD distance between their latent spaces.

Figure 5.13 details the pipeline of the LMMD-FS proposed method.

5.7 Evaluation of the selected features

As used in chapter 03, to evaluate the quality of the selected features the Redundancy Rate (RED) metric is used[56] [48].

To evaluate visually the quality of the selected features it is possible to use a non-linear dimensionality reduction method that projects the tumor samples characterized by the selected p features in a two-dimensional representation used only for visualization purposes.

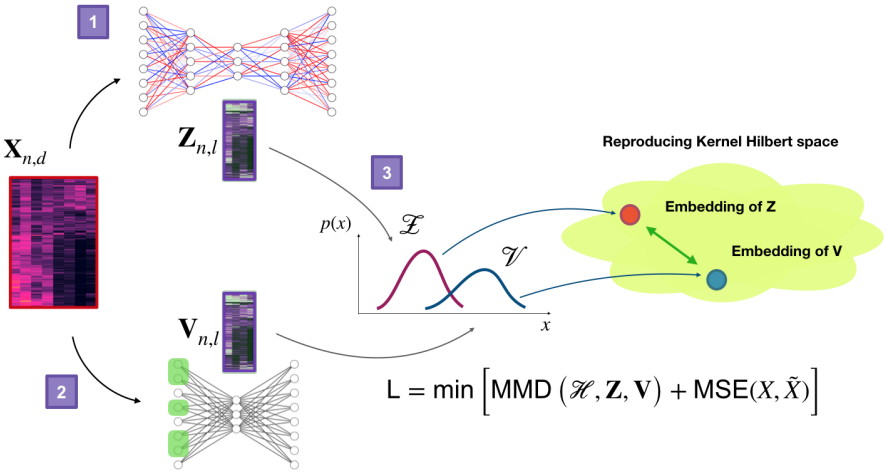

Figure 5.13: A deep regularized autoencoder (1) and a vanilla autoencoder (2) are simultaneously trained from the input data. The loss function (3) considers the reconstruction loss but also the MMD distance between the latent space of the two autoencoders. Features are selected based on a ranking computed with the first layer of the vanilla autoencoder and a penalization term.

Then it is possible to evaluate how the tumor samples are distributed in the new feature subset. The t-distributed stochastic neighbor embedding (t-SNE) [133] is used for this purposes. Finally a third evaluation approach is by the clustering performance on the selected features from a biomedical perspective via the Rand Index as explained in section 5.3.2.

5.8 Experiments

The experimental section first presents the real cancer genomics datasets that are used, then the corresponding preprocessing treatment done on each one and finally the results.

5.8.1 Datasets

To evaluate the two proposed methods three tumor datasets are used: Lung, Renal and Brain cancer. The reason by which these datasets are used in this chapter is based on the fact that each tumor type on each dataset is composed by two tumor subtypes classes used as ground truth for tumor subtype discovery. All datasets have been downloaded from the International Cancer Genome Consortium (ICGC) data portal.
The Lung Cancer dataset is composed of the projects LUSC-US (Squamus cell subtype) and LUAD-US (Adenocarcinoma subtype) with 478 and 428 tumor samples respectively. The Brain cancer dataset is composed by the projects GBM-US (Glioblastoma) and LGG-US (Lower Grade Glioma) with 159 and 439 tumor samples respectively. The Renal cancer data is composed by the projects KIRP-US (Papillary) and KIRC-US (Clear cell) with 222 and

518 tumor samples respectively. The data matrix for each dataset is X_{nd_0} with n tumor samples and characterized initially by $d_0 = 17640$ protein coding genes. The objective is to select a subset of genes $p << d_0$ to reduce the dimensionality and retain the discriminant information about cancer subtype.

Cancer Type	Tumor Subtype	Project ID	Patients (n)
Lung Dataset	Squamus cell	LUSC-US	478
	Adenocarcinoma	LUAD-US	428
Renal Dataset	Papillary	KIRP-US	222
	Clear cell	KIRC-US	518
Brain Dataset	Lower Grade Glioma	LGG-US	439
	Glioblastoma	GBM-US	159

Table 5.3: Number of patients by tumor type and subtype.

Table 5.3 details the three datasets associated to each cancer type. Each dataset is composed of two classes associated to the tumor subtype.

5.8.2 Pre-processing

To estimate statistically the performance of the proposed methods ten independent times 80% of the samples have been randomly selected from the input dataset to train the proposed methods and select the gene features. For both autoencoder training 80% of the randomly selected samples are used and the remaining 20% used for validation of the neural network. Each initial feature has been min-max scaled between 0 and 1 [134]. Then an univariate filter is applied to reduce the initial number of features from $d_0 = 17640$ to $d = 8820$ by ranking the features by variance and only keeping the 50% best ranked. Univariate filter is used to discard low variance features and perform an initial reduction of the high dimensional space. Nevertheless, the subset of d variables preserved remains large and the sample to feature ratio for the three datasets is $(n/d) < 0.11$ thus the proposed feature selection method is applied at this point.

5.8.3 Results

Experiments on each dataset have been conducted ten times by randomly selecting 80% of the data. At each random iteration and in the following order data pre-processing, feature selection and clustering are performed. Then the evaluation by RED and Rand Index is averaged among all random iterations. A set of different number of selected p features is used where $p = [10, 20, 30, 40, 50]$. A set of k clusters are obtained where $k = [2, 3, 4, 5]$.
The first evaluation to be done is the RED score for each subset of selected features p by each method on each dataset.

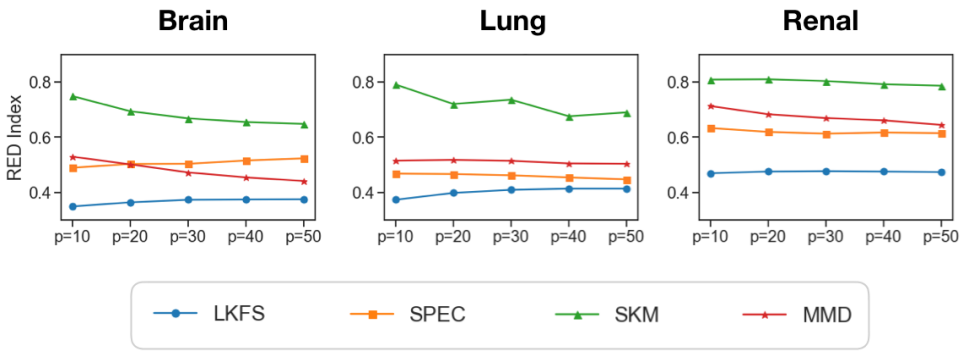

Figure 5.14: RED score on the selected features of each method.

Figure 5.14 shows the results of the RED score. It is observed that the proposed LKFS has the lowest RED score in all the experiments followed by the SPEC method and the LMMD-FS while SKM has the highest RED. This evidences that the features selected by the LKFS have the lowest redundancy which is a desired result since these are more informative and less redundant. This LKFS result may be caused by the selection of non-redundant feature-wise kernels in the MKL task therefore the selected features may be not redundant as well.

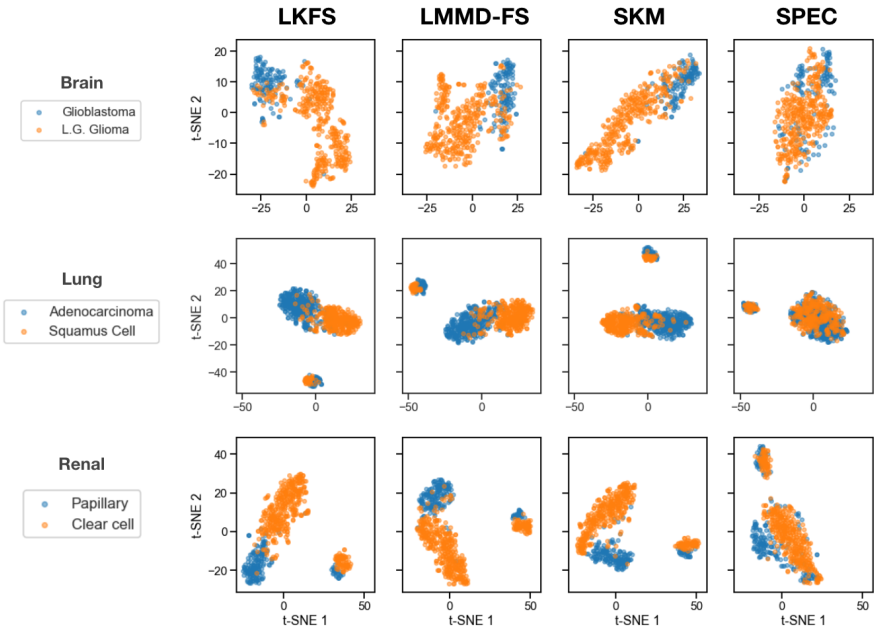

Figure 5.15: Visualization of the distribution of samples on the selected features using each method with $p = 50$ on each cancer dataset.

Figure 5.15 shows a two dimension t-SNE scatter plot of each dataset by each method

with $p = 50$. For the Brain, Lung and Renal datasets LKFS shows how clusters tend to group tumors of the same subtype together in two main clusters. The LMMD-FS shows a similar trend. The SKM tend to polarize different tumor subtypes within the same cluster structure and finally SPEC fails to separate effectively the two tumor subtypes of each dataset.

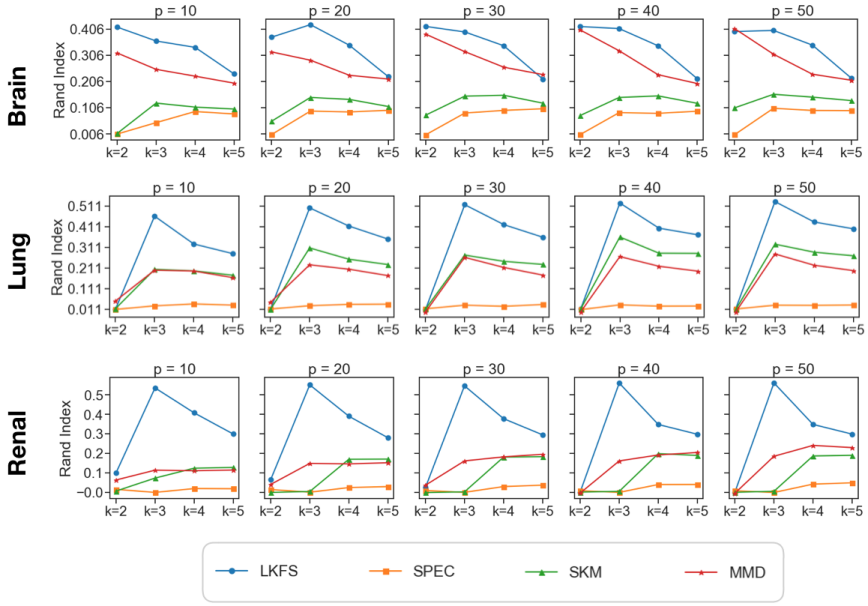

Figure 5.16: Rand index performance of each feature selection method using different values of selected features p on each dataset for different values of clusters k.

Then K-means is applied on the selected features and the quality of clusters measured by the Rand-Index. Figure 5.16 shows the Rand-Index for different number of k clusters on different number of p selected features for each dataset and for each method. In the Brain dataset the LKFS method outperforms all the methods followed by the LMMD-FS while the SPEC and SKM methods have the lower values of Rand Index. for every value of p and k. In both the Renal and Lung datasets the LKFS outperforms the benchmark methods from $k = 3$ to $k = 5$ and the LMMD-FS achieves a similar performance with SKM. For these two datasets the four methods have a considerably low rand-index with $k = 2$ since in every case a small and isolated cluster composed of the two subtypes affects the performance as can be observed in Figure 5.15.

To better understand the LKFS method and why it outperforms the rest of the benchmark methods the kernel matrices K_z and K_μ are visualized. Figure 5.17 shows the resulting K_z kernel matrices obtained from the latent space of deep regularized autoencoder on each dataset and compared with the K_μ obtained from the MKL task.

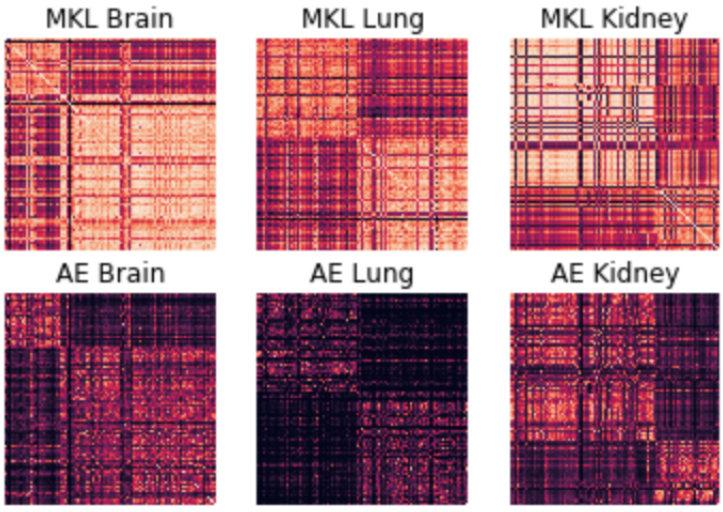

Figure 5.17

Figure 5.17 shows how both matrices reveal similar structures evidencing that the MKL task well captures the latent structure learnt by the deep regularized autoencoder and consequently the informative features for the clustering task.
Additionally the clustering performance is evaluated on the latent space of the deep regularized autoencoder and on the selected features of the LKFS method. Figure 5.18 shows how both clustering performances are almost equal for different number of clusters proving that the structure information is preserved in the selected features by the MKL task.

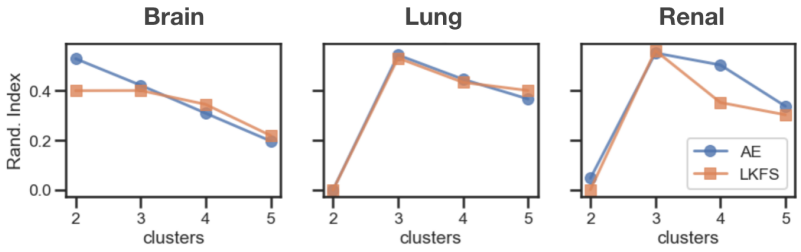

Figure 5.18

5.9 Discussion

This chapter proposes the LKFS and the LMMD-FS methods where both use an unsupervised approach to select a reduced gene subset from more than 8.000 genes to less than 50 in three different types of cancer datasets with the objective of improving the clustering performance. The selection of the LKFS features is done by improving the alignment between a kernel

matrix K_z obtained from the latent space $\mathcal{Z}$ learned via an autoencoder and by the resulting one K_μ after multiple kernel learning on feature-wise kernels. The selection of the LMMD-FS is done by minimizing the MMD discrepancy between the latent space of two autoencoders and building a feature ranking based on the L1 norm. Both proposed approaches are based on building a target representation of the input data using the latent space of deep regularized autoencoder. This target representation has a reduced level of noise and is low dimensional. LKFS and LMMD-FS select only the features that align the most by kernel alignment and reduces the MMD-distance to the target representation respectively. The target representation can be interpreted as a prior distribution that guides the selection process.
To measure the quality of the unsupervised feature selection process the Redundancy score RED is computed on the selected features. In addition a K-means clustering method is applied as a downstream task and the cluster performance is evaluated by the homogeneity of ground truth labels across clusters by the Rand Index.
From the experimental results it is observed that LKFS outperforms the LMMD-FS, the SKM and SPEC methods by Rand Index and RED score for almost all the datasets and number of selected features. This suggest that LKFS is able to selects low redundancy features from high dimensional input space that contributes to find well defined clusters composed mainly from one tumor subtype. Additionally the LMMD-FS method has a similar behaviour of SKM method in almost every dataset for any number of clusters with the exception of the Brain dataset where the Rand Index obtained is the second highest. In contrast, SKM and SPEC select features with higher redundancy and relatively small Rand Index which do not contribute enough to build separated clusters neither to group samples from the same tumor subtype together.
One of the advantages of both LKFS and LMMD-FS is the target representation of the Autoencoder. This representation captures the salient features of the input dataset and then by MKL or MMD the selected features will capture approximately the same data structure. It is observed how LKFS is a more complex model in comparison with the LMMD-FS since the first involves thousands of initial feature-wise kernel matrices showing that the increasing in complexity improves the quality and selection performance. On the other side, despite the performance of the LMMD-FS is not the highest it is based on a simpler approach since it just involve two autoencoder models instead of thousands of kernel matrices therefore a trade-off in complexity has to be considered. Finally, LKFS and LMMD-FS methods provide two outputs. The first is the subset of selected gene features which is considerably reduced in comparison with the original feature set and helps for biological interpretation. The second one is the latent space provided by the autoencoder. The latent space serves not only as a target representation for the feature selection process but also as a tool for data exploration and analysis since it can summarizes in a lower dimensional space the salient features of the original data.
One limitation of LKFS and LMMD-FS relies in training two models, the deep regularized autoencoder and the MKL or Vanilla Autoencoder respectively. Then the feature selection approaches are conditioned by the architecture and quality of the deep regularized autoencoder.

5.10 Conclusions

This chapter proposes two unsupervised feature selection methods named LKFS and LMMD-FS capable to select a considerably reduced subset of meaningful and low redundant features from high dimensional gene expression data. The LKFS method is based on autoencoder and multiple kernel approaches. The LMMD-FS is based on two autoencoders. Experimental results show that the proposed LKFS method outperforms two benchmark unsupervised feature selection algorithms by analyzing the quality of the clusters built on the selected features. Additionally it is observed that the LMMD-FS achieves similar results to the SKM method. The results suggest that the higher complexity based on multiple kernels of the LKFS is an advantage in achieving the highest clustering and selection performance while the LMMD-FS with a simpler architecture equals the SKM method. Both methods reveals the importance of using a latent space to guide the feature selection process by two different approaches: Kernel Alignment and Maximum Mean Discrepancy. The proposed methods have been evaluated on tumor gene expression datasets from Lung, Renal and Brain Cancer patients and select features that help to identify tumor subtypes without any supervised approach. For this reason LKFS and LMMD-FS are useful models for pattern recognition and data mining in a variety of cancer types and high dimensional biological applications. The next chapter extends the LKFS method to a latent space learned from multi-omic data such as genomics and transcriptomics simultaneously.

5.10.1 Scientific Production of this chapter

The content of this chapter has been presented in:

- LKFS method has been published at the IEEE CBSB 2020 international conference [129].
- LKFS method won the best Poster presentation at Litoral AI CONICET 2019
- LKFS method won the best Poster presentation at UTN Doctoral event 2019
- LKFS Poster presented at Khiphu AI Latin American Meeting on Artificial Intelligence
- LKFS Poster presented at the Functional Inference and Machine Intelligence Workshop 2020
- LKFS Poster presented at the Ibero-american conference on Bioinformatics 2019

Chapter 6

Unsupervised feature selection guided by multi-omics latent space

6.1 Introduction

Molecular data from tumor profiles can be divided in multiple groups of random variables associated to measurements done in specific contexts. Each of these groups is a specific source of variability measured in different molecular contexts. The multi modal scenario is described by sources of variability known as *omics* where each *omic* is associated to a particular type of molecular features: genome, transcriptome, proteome and metabolome [5]. Each omic is composed of macromolecules that encode biological information about cell function by following the central biology dogma and has a different impact in biological regulation within a cell.

To get data from the different omic environments different technologies are used. For instance, the high-throughput next-generation sequencing (NGS) technology mentioned in chapter 01 is used to get the genomics (DNA mutations) and Transcriptomics (RNA concentration) while the Mass Spectrometry technology [135] is used to get data from Proteomics (protein concentration).

Despite the cost of sequencing different omics is decreasing globally the complexity to obtain multiple omics for each cancer sample is still a limitation caused by equipment or lab capacity for this reason some experiments get data from only a single omic, commonly *Transcriptomics* and a lack of the rest of the omics. In this thesis from chapter 02 to chapter 04 the *transcriptomic* or gene expression *omic* has been used since it is a useful source of information to determine the state of a given cell and is one of the most used omics in single cell analysis and cancer genomics [136] [137] [138].

Neverheless if a set of tumor profiles is described by multiple omics instead of learning from a single one it is feasible to propose methods that learn simultaneously from different omics in order to improve a specific learning task. It is assumed that each omic provides specific information about the data distribution, thus learning from multiple ones may explain better the data manifold.

Since single omic data is high dimensional multi-omic data is even higher dimensional, thus this chapter studies and analyzes if it is possible to learn a low dimensional latent space when multiple omics are used as input data and fusion them in a single representation. To learn a single low dimensional latent space $\mathcal{Z}$ from multi-omic data the *Multi-Modal Learning* approach can be used by relating information from multiple sources and as consequence improve a downstream learning task in the latent space [139] with the assumption that each omic may explain some particular structure of the data by their own. In this work the downstream learning task is clustering.

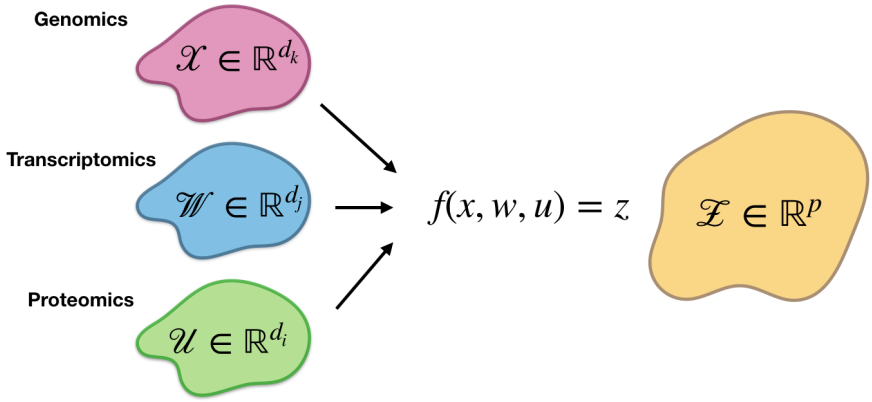

Figure 6.1: Diagram of the multi-omic fusion problem. From multiple omics $\mathcal{X}$,$\mathcal{W}$,$\mathcal{U}$ a low dimensional representation $\mathcal{Z}$ is learned.

The multi-modal learning approach and multi-omic data fusion can be implemented by different strategies. It can be done by using Autoencoders [139][140] or by using kernel methods, more precisely combination of kernels and kernel-PCA [43].

Once the latent space is learned from the multi-omic input data, an unsupervised feature selection task can be done on the transcriptomic input features as done in chapter 03 with the Latent Kernel Feature Selection Method (LKFS) [129] where the goal is to guide via a latent space the selection of a subset of the input features that keeps the data structure necessary to find clusters or communities of tumors with common clinical attributes. The main difference in this work with respect to the LKFS method is that the latent space learned represents a data structure coming from multiple omics instead of a single one. If the latent space learned from multi-omic data has a better structure than a latent space learned from single omic data based on the clustering performance then the unsupervised feature selection task from LKFS can be improved if multiple omics are used for latent space learning. In this work two omics are used in the experimental section: Simple Somatic Mutations (Genomics) and RNA-seq (Transcriptomics). To understand the impact of doing latent space learning and unsupervised feature selection two scenarios are evaluated: single omics where all the samples are just described only by the transcriptomic features and multi-

omics where all samples are characterized by both simultaneusly transcriptomics and genomics.

This chapter is organized as follows: section II presents the related work and benchmark scenarios of multi-omic data fusion and dimension reduction, section III defines the concept of multi-omic data fusion with Autoencoders and Kernel Methods, section IV describes the synthetic and real datasets used and section V the experimental results.

6.2 Related work

The multi-omic data is represented by multiple groups of biomolecular features such as DNA mutations (genomics), RNA Transcript concentration (Transcriptomics) or Protein concentration (Proteomics). As explained in chapter 01 of this thesis, each of these *omics* contains multi-dimensional variability about different aspects of a cell. The genomic group of features explain the alteration in single bases of DNA, known as mutation, which occur within each protein coding gene. The Transcriptomic group of features involves the concentration within the cell of RNA transcripts associated to a gene and represent if a given gene is expressed and active or not. Finally from the RNA transcripts Proteins are built. Proteins define their function based on molecule structure, therefore the concentration of a given protein within the cell indicates the cell function. The multiple-omic features together are a useful source of data that may reflect the state of a cell and help to estimate the phenotype such as tumor subtype [106] [141].
From a machine learning perspective multi-omic data fusion and dimension reduction has been developed and applied using both Neural Networks and Kernel Methods.
Autoencoders have been used on a wide range of applications in cancer informatics. One application is its use on a single cancer type, such as liver cancer, while combining multi-omics data from tumor profiles [106] to learn a latent space and identify new cancer subtypes. A similar case has been proposed for breast cancer to discover subtypes using transcriptomics data [110]. A newer version of AE, the Variational Auto-encoder, has been used to learn a latent space to improve the classification of known subtypes of lung cancer using DNA methylation data [112]. Moreover, instead of learning a latent space from a single type of cancer, a pan-cancer study based on transcriptomics data from The Cancer Genome Atlas (TCGA) [3] using Variational Auto-encoders evidenced a big potential for the use of autoencoders to learn reduced latent space while keeping biological insights [142]. Another work with gene expression data from TCGA applied standard autoencoders and Gene Supersets, which are a priori defined gene sets that retain biological meaning in the latent space [143]. On the other hand, a network and graph theory analysis has been done for pan-cancer mutational data to detect communities of tumors [144] using the co-occurrence of mutations as connections. A recent work maps mutated genes instead of the tumor samples to a lower dimension using deep learning techniques to learn a distributed representation of molecules [145].
Kernel methods have been also used for the multi-omic data integration problem. Multiple Kernel Learning has been used to merge multi-omic data in breast cancer samples for

unsupervised clustering tasks [146]. Additionally Kernel Methods have been used to merge and fusion multi-omic data from breast cancer tumor profiles in supervised learning tasks for subtype classification [147]. Moreover, Kernel Method have been used to integrate Transcriptomics, Genomics and Methilation data from pancancer tumor profiles to predict patient survival as a supervised regression task [5].
By reviewing the bibliography, it is clear that data from different omics sources require models to simplify the original context and reflect emerging patterns. Autoencoders and Kernel Methods have shown great adaptability to biological data and are extremely useful for reducing dimensionality.
In the next section the multi-omic fusion task for dimension reduction and latent space learning by using Kernel Methods and Autoencoders is explained.

6.3 Multi-omic fusion

Multi-omic fusion methods aims to learn a low dimensional representation of tumor samples characterized by multiple omics. The low dimensional representation captures the variability from the multiple input omics in a single representation $\mathcal{Z}$. This means that if the multi-omic fusion method is well implemented it is expected to capture data structure coming from all the input omics thus the multi-omic fusion may explain structure that a single omic can not alone.
As detailed in the related work section, in this chapter two alternatives to fusion multi-omic data are studied. The first one is via multi-modal autoencoders where multi-omic data is used to learn a single low dimensional latent space $\mathcal{Z}$ via an Encoder function and reconstructed from $\mathcal{Z}$ to $\mathcal{X}$ via a Decoder function. The second alternative used in this work is the combination of kernels each one used on each input omic and then merged in a single kernel which corresponding representation captures variability and structure from all omics.
In this work both strategies are used to learn a joint representation from Simple Somatic Mutation (Genomics) and Gene Expression Sequence (Transcriptomics) data of tumor profiles. The tumor profiles are described by random variables coming from both omics where d_x and d_u are the dimensions that defines the spaces $\mathcal{X} \in \mathbb{R}^{d_x}$ and $\mathcal{U} \in \mathbb{R}^{d_u}$ respectively. For this reason a tumor sample can be described by a single omic if there is only one of those available or by both in case the two are available. Therefore a set X of samples characterized by a single omic $\mathcal{X}$ samples is defined as

$$X = \{x_1, \ldots, x_n\}$$

or characterized only by $\mathcal{U}$ defined as

$$U = \{u_1, \ldots, u_n\}$$

and on the other side a multi-omic sample set is defined when the two omics are available as

$$M_{x,u} = \{m(x_1, u_1), \ldots, m(x_n, u_n)]\}$$

where n corresponds to the number of samples, x the samples described only by the first omic, u the samples described only by the second omic and m the ones described simultaneously by the two omics. In the experimental section the dimensionality reduction methods used are evaluated on single omic and multi-omic datasets in order to analyze how learning from multi-omic data improve a downstream task like clustering.
The following subsections explain the methodology used to fusion multi-omic data in a single representation $\mathcal{Z}$ by Autoencoders and Kernel Methods.

6.3.1 Multi-omic fusion with Autoencoders

This section defines the multi-modal autoencoder.

The standard autoencoder model is described previously in chapter 05 of this thesis where the input and the output are one single modality. In this chapter the standard autoencoder model is adapted to deal with multi-modal data such as multi-omic tumor profiles. For this reason the multi-omic autoencoder is composed by two encoders f_x and f_u contatenated by a fully connected layer f_z such as

$$f_m(X, U) = f_z(f_x(x), f_u(u)) = z \tag{6.1}$$

and the decoder side composed by two decoders q_x and q_u

$$q_m(z) = q(q_x(z), q_u(z)) = (X, U) \tag{6.2}$$

In the proposed multi-omic autoencoder the encoder function has two inputs that take both modalities X and U of each sample $M(x, u)$ and map them to their corresponding low dimensional representation z_x and z_u. Then both latent representations are merged in a feed forward fully connected layer into the low dimensional latent space $\mathcal{Z}$ where the structure of both modalities is merged to a single representation. The decoder function has as input the latent samples from $\mathcal{Z}$ and reconstruct the two modalities $\hat{X}$ and $\hat{U}$ as output as detailed in figure 6.2.

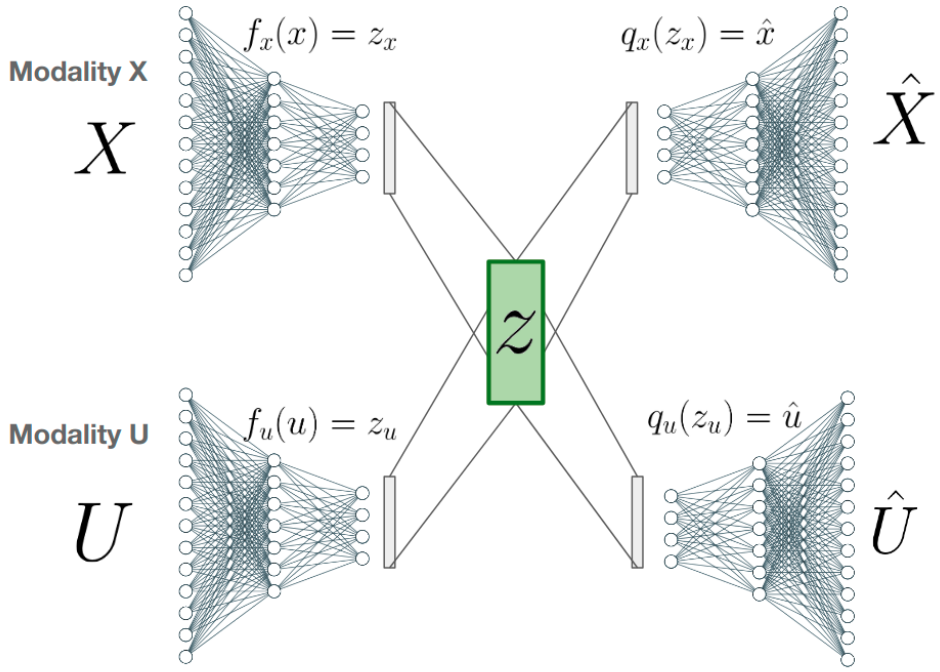

Figure 6.2: Proposed architecture for the multi-modal autoencoder.

Then the loss function of the multi-modal autoencoder must track the reconstruction quality of both modalities for this reason it computes the divergence between the input modalities and the reconstructed ones as

$$L_{(x,u)}\left[(x,u),(\hat{x},\hat{u})\right] = L_x + L_u = L(x,\hat{x}) + L(u,\hat{u})$$

thus the optimization process will consider two omics simultaneously instead of one.

6.3.2 Multi-omic fusion with Kernel Methods

To benchmark the multi-omic autoencoder performance a multi-omic integration task with kernel methods is presented. A brief summary of Kernel Functions is detailed as described in chapter 03. Kernel methods are based in a nonlinear *feature map* ϕ from the input space $\mathcal{X}$ to a high dimensional space known as *feature space embedding* $\mathcal{H}$. Given $\mathcal{X} \subseteq \mathbb{R}^d$ a d-dimensional space , where ϕ is implicitly defined by choosing a kernel function k that implement the dot product in the feature space.

As explained in chapters 3 and 5 of this thesis, Kernels can be combined by different operations. A key operation for multi-omic data fusion is the sum of kernels. Two kernels k_x and k_u can be linearly combined and the result is a valid kernel such as

$$k_z\left([\mathbf{x}_1,\mathbf{u}_1],[\mathbf{x}_2,\mathbf{u}_1]\right) = \quad k_x\left(\mathbf{x}_1,\mathbf{x}_1\right) + k_u\left(\mathbf{u}_2,\mathbf{u}_2\right) \tag{6.3}$$

The sum operation can be used to mix kernels associated to different omics and the resulting k_z is a kernel containing the pair-wise similarity between samples from the two involved omics. Then with the resulting kernel from the mixture of the two omics the Kernel-PCA method can be used to learn a low dimensional non-linear representation of the input multi-omic data.

As detailed previously by doing multi-omic data integration with kernel-PCA the kernel used is the one obtained from the linear combination of the omic-wise kernels as detailed in figure 6.3

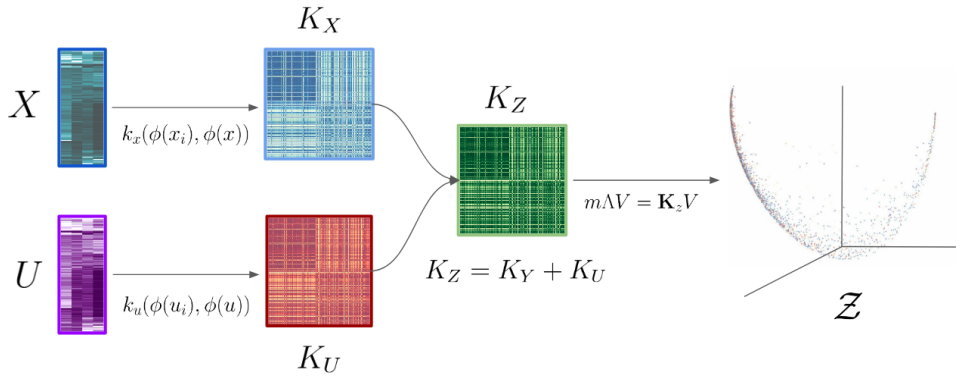

Figure 6.3: Proposed pipeline for the multi-modal kernel dimension reduction.

6.3.3 Evaluation of multi-omic latent space

In this work the evaluation of the latent space obtained from the multi-omic data sources is validated by an unsupervised *downstream* task. This means that a clustering method is applied to group the multi-omic tumor samples in the learned low dimensional latent space after model training. If the latent space correctly captures the inner structure of the multi-omic tumor profiles and the multi-modal dimensionality reduction method filters the noise from input data then the clustering performance has to be improved in the latent space when compared to the input space.
To evaluate the quality of the learned latent space $\mathcal{Z}$ the K-means clustering method [103] is applied on the tumor samples z projected on the latent space as $z \in \mathcal{R}^p$ where p is the latent dimension with $p < d$. A k number of clusters is determined as an external hyper-parameter and each tumor sample z_i is assigned to a cluster label c_i. As in chapter 05 to evaluate the results of the clustering method the Adjusted Rand Index is used [108].

6.4 Proposed Method: Latent Multi-omic guided feature selection

The chapter 05 of this thesis proposes the Latent Kernel Feature Selection (LKFS) method designed to perform unsupervised feature selection on single omic data. In this section first the LKFS method is briefly described and then extended to multi-omic data.
The LKFS method is designed to first learn a latent space $z = \phi(x)$ where the mapping function ϕ is trained by an unsupervised dimension reduction method and x is the single-omic training data. A kernel $k_{AE}(z_i, z_j)$ is learned on the latent space and used to build a kernel matrix K_{AE} that details the pair-wise similarity measurements between samples lying on $\mathcal{Z}$. Then from a set D of d feature-wise kernel matrices K_i a Multiple Kernel Learning approach determines a linear combination $K_\mu = \sum \mu_i K_i$ with the objective to maximize the alignment $A(K_{AE}$. The solution to maximize the objective function $A(K_{AE}, K_\mu)$ may be sparse so that many terms are $\mu_i = 0$.
To extend LKFS to multi-omic data a dataset $M = [m(x_1, u_1), ..., m(x_n, u_n)]$ is used to train a multi-modal autoencoder instead of the single modality autoencoder. A latent space $\mathcal{Z}$ is obtained and used as target kernel to guide the feature selection task on transcriptomic data with the Multiple Kernel Learning Step.

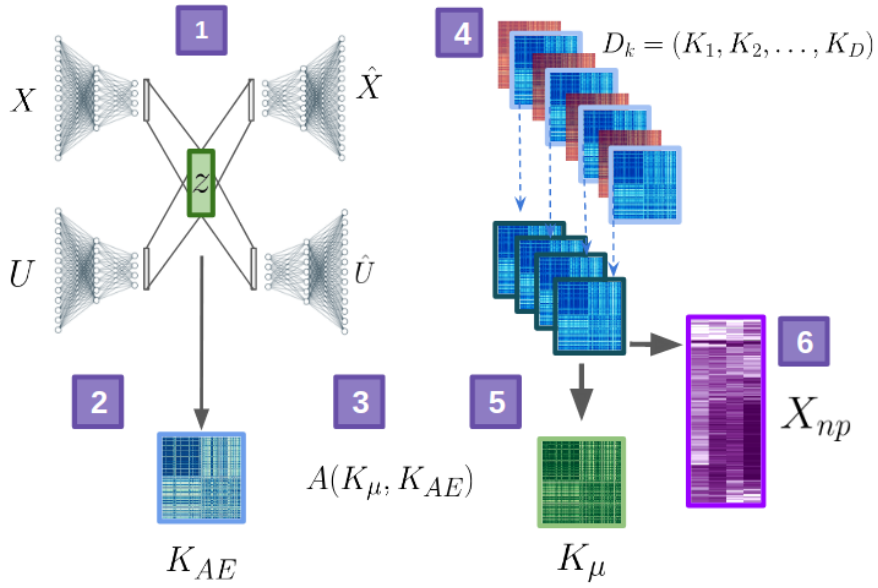

Figure 6.4: Proposed pipeline for the unsupervised multi-omic pan-cancer feature selection. First a latent space is learned and a target kernel matrix K_{AE} built. Then a set D of feature wise kernels is used to solve a multiple kernel learning problem where the resulting kernel K_μ improves the alignment $A(K_{AE}, K_\mu)$. Finally a subset of p features is obtained.

Features are still being selected from a single omic, in this case transomics. Nevertheless the latent space is learned from multi-omics. This may cause that the selected features

from transcriptomics will respond to the structure obtained from multi-omics and eventually changing the solution when compared to the single-omic feature selection task. If a latent space from multi-omic input data reveals better structure of data distribution than a single-omic latent space then the selected features may also improve the structure of the resulting distribution and the clustering performance.
Figure 6.4 shows the pipeline of the proposed unsupervised multi-omic feature selection method.

6.5 Datasets

To evaluate the proposed multi-omic dimension redution and multi-omic integration approach two datasets are used. The first one is a Synthetic Dataset known as *spiral dataset* and the second one is a real pan-cancer multi-omic dataset.
Then for the feature selection task the pan-cancer multi-omic dataset is used.

6.5.1 Synthetic dataset

The first dataset serves as a synthetic toy example known as *spiral dataset* where the data distribution is known. It is composed of two classes, each class has $n = 100$ samples. The original intrinsic dimension of the dataset is $d = 2$ with features x_a and x_b. Both classes are distributed in a two dimensional spiral manifold as detailed in the upper left box of figure 6.8. To make the dataset high dimensional 98 noise dimensions are added for each of the two original ones using a uniform distribution $x_i \sim U(0, 1)$. The final dimension is $d = 200$ composed by two groups of $d_a = 100$ and $d_b = 100$ features, each group redundant to x_a and x_b respectively. To make the dataset multi-modal each group of features d_a and d_b is considered as a modality or *omic*, thus the spiral dataset has two modalities.

d_a	d_b	Total dimension d	Sample size n
100	100	200	200

Table 6.1: Synthetic spiral dataset size.

Table 6.1 details the size of the synthetic dataset. The proposed dataset has the property that each modality can not explain the spiral manifold by its own and the only way to capture the ground truth manifold is by learning from both modalities simultaneously.

6.5.2 Pancancer dataset

To evaluate the proposed methods on real data a *multi-omic pancancer* dataset is used. From the International Cancer Genome Consortium (ICGC) data portal [16] a total of $n = 3736$ samples of eight cancer subtypes have been downloaded. The subtypes are Breast Cancer

(BRCA-US), Head and Neck Squamous Cell Carcinoma (HNSC-US), Kidney Renal Clear Cell Carcinoma (KIRC-US), Brain Lower Grade Glioma (LGG-US), Lung Adenocarcinoma (LUAD-US), Lung Squamous Cell Carcinoma (LUSC-US) and Skin Cutaneous melanoma (SKCM-US). Similarly to the example of Figure 1 each tumor profile is characterized by two modalities each one associated to a specific *omic*: Genomics and Transcriptomics. Both modalities are composed by gene features. The difference between them is the biological information encoded within the genes. The Genomic modality represents the number of Somatic Point Mutations observed in the DNA sequence within each gene. On the other side the Transcriptomic layer encodes the level of expressed RNA molecule (transcripts) associated to each gene.

Genomics (d_{ssm})	Transcriptomics (d_{exp})	Total dimension d	Sample size n
9025	7498	16523	3736

Table 6.2: Pan-cancer multi-omic dataset size.

Table 6.2 details the dimensionality of each *omic* and the sample size of the pancancer dataset. In total there are 8 ground truth classes each one associated to a single tumor subtype. Additionally the number of samples per subtype is detailed in figure 6.5.

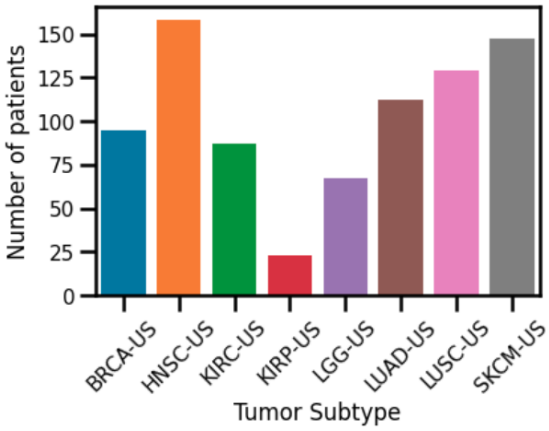

Figure 6.5: Number of samples per tumor subtype.

As detailed in the materials and methods section the objective is to learn a single and low dimensional representation from the multi-omic input data and evaluate it by an unsupervised downstream task. Therefore the tumor subtype labels are not used during the learning and training process and only considered during the final evaluation step.

6.6 Experimental results

This section is devoted to show experimental results on the synthetic and real cancer datasets. Synthetic dataset is used to evaluate if both the multi-modal autoencoder and the combination of kernels can capture the ground truth spiral manifold by learning from both omics. The pancancer dataset is used to evaluate if the proposed methods can learn a low dimensional latent space that allows a accurate clustering of tumor subtypes.

6.6.1 Results on synthetic dataset

As explained in the previous section, the objective of the synthetic dataset experiment is to capture the two dimensional spiral manifold from the multimodal input data. The multimodal autoencoder used has an encoder and decoder functions for each of the modalities. From input to output after the two encoder functions both are merged in a single representation followed by two decoder functions to reconstruct both modalities. Figure 6 details the proposed architecture for the multimodal autoencoder with the current number of neurons per layer.

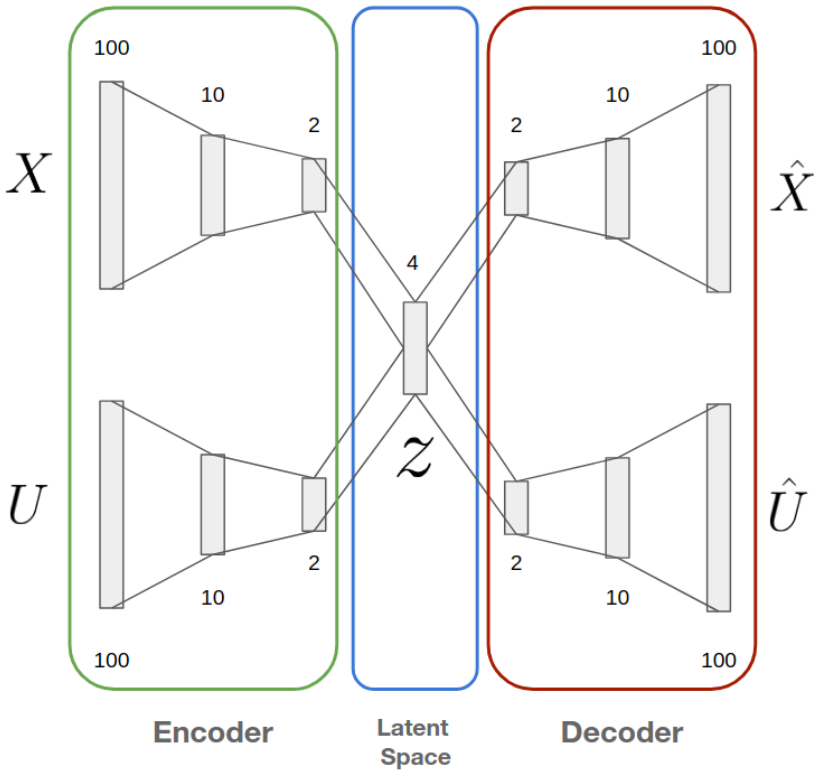

Figure 6.6: Multi-modal autoencoder proposed architecture for the toy example case.

On the other side the kernel combination method is used to fusion a kernel corresponding to each modality k_x and k_u in a single multi-modal kernel $k_z = k_x + k_u$. As explained in

the Materials and Methods section the resulting kernel k_z is used to learn a low dimensional latent representation by kernel-PCA method.
Figure 6.7 shows a plot of the kernel matrices of each single modality and the combined ones. Figure 6.8 shows a two dimensional scatter plot of each method for each single modality and the resulting multimodal one. The upper row of plots shows the ground truth distribution (upper left) and the multi-modal representation obtained by the autoencoder (upper center) and the kernel combination (upper right). The center and lower rows shows the single modality distribution for each method.

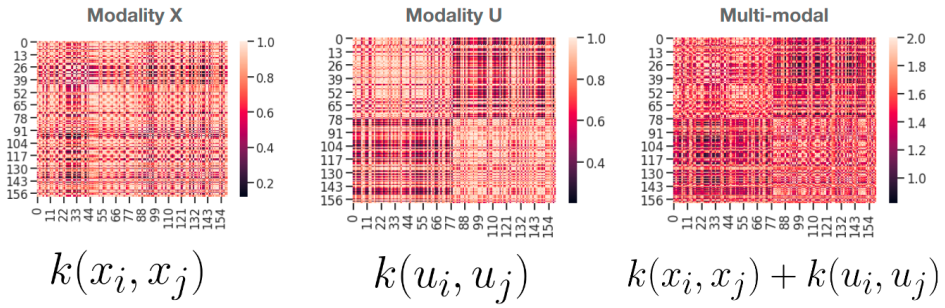

Figure 6.7: Kernel matrices corresponding to modality X (left), modality U (center) and to the multimodal representation (right).

Figure 8 shows two dimensional representation of the obtained latent distributions.

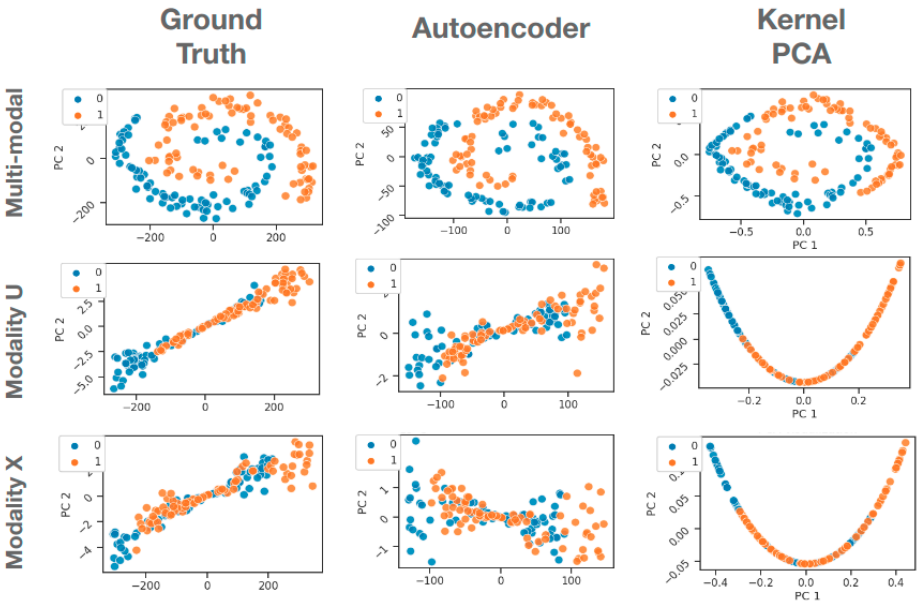

Figure 6.8: Visualization of the distribution of samples on single and multi modal approaches using a PCA projection of the resulting latent space in two dimensions.

It is observed how each modality by their own cannot explain the spiral manifold. By both the autoencoder and kernel combination methods the spiral manifold emerges as a representation learned from both modalities. Despite the spiral dataset is low dimensional and relatively easy to solve, it serves as a visual evaluation and interpretation of how the multi-modal dimension reduction methods can obtain the desired target distribution.

6.6.2 Pancancer dataset

As described in the materials and methods section to deal with the pancancer multi-omic input data a specific autoencoder architecture is proposed.

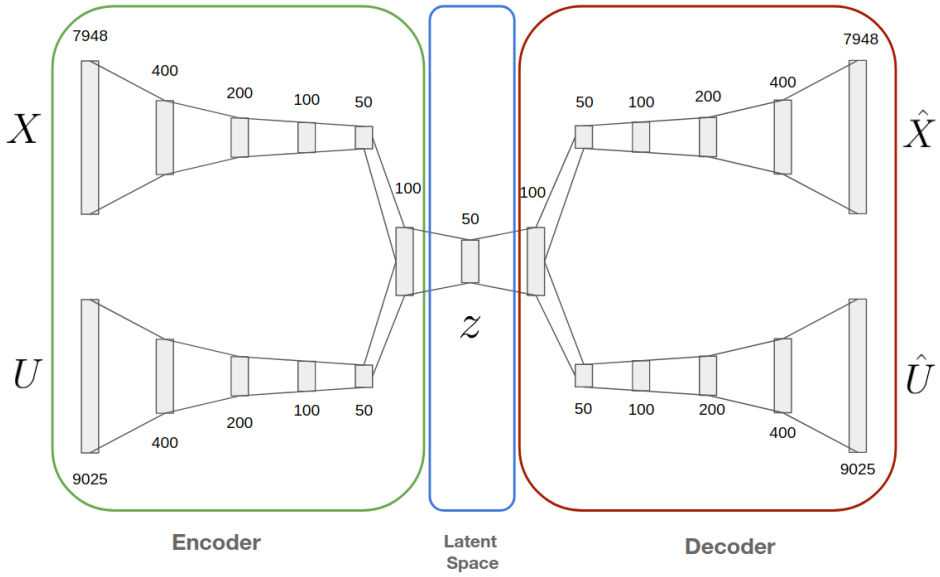

Figure 6.9: Proposed architecture for the Multi-modal autoencoder. The number of neurons is detailed on each layer.

Similarly to the synthetic dataset experiment the model used in the pancancer one is composed of two encoder and two decoder functions as detailed in figure 5.9. The encoders are merged downstream in a single representation followed by a split into the two decoder functions. The input dimension is $d_{\text{multi}} = d_{\text{ssm}} + d_{\text{exp}} = 16523$ while the latent dimension of z is $p = 50$ therefore the compression is $\frac{16523}{50} = 330$ times. The multi-omic samples are processed by the encoder network and mapped to the low dimensional latent space $\mathcal{Z}_{\text{ae}}$.
On the other side the kernel combination method is also used to learn a low dimensional representation of the input multi-omic pancancer data. To achieve that objective two gaussian kernels are learned from the training data. One kernel corresponding to the Genomic modality as k_{ssm} and one corresponding to the transcriptomic modality as k_{exp}. Then both kernels are linearly combined as expressed in equation 6.4.

$$k_{\text{multi}}\left([\mathbf{x}_i, \mathbf{x}_j], [(\mathbf{u}_i, \mathbf{u}_j)]\right) = \quad k_{\text{ssm}}\left(\mathbf{x}_i, \mathbf{x}_j\right) + k_{\text{exp}}\left(\mathbf{u}_i, \mathbf{u}_j\right) \tag{6.4}$$

The sigma parameter σ_{exp} and σ_{ssm} corresponding to k_{exp} and k_{ssm} respectively are computed by the median pairwise distance between samples on each omic. Once the k_{multi} is computed on the entire dataset an $n \times n$ kernel matrix is obtained as detailed in figure 6.10 where the genomics, transcriptomics and multiomics matrices are detailed.

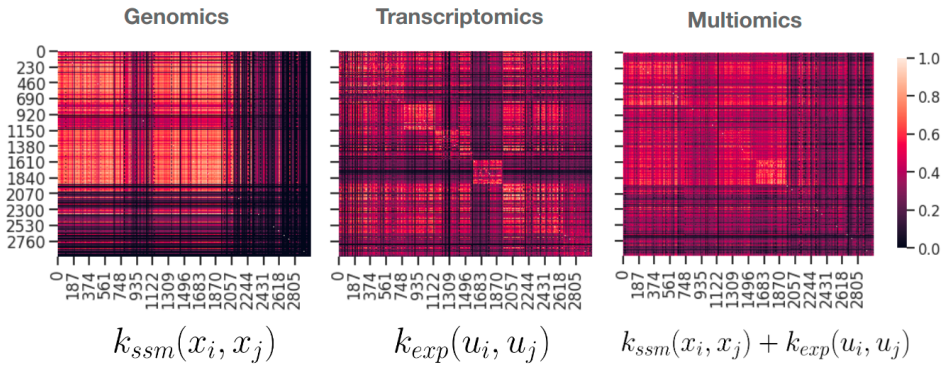

Figure 6.10: Visualization of the distribution of samples on the selected features.

Visually it is observed that each matrix captures different similarities between tumor profiles thus the distribution captured by each kernel is different and the final multi omic kernel may capture the combination of both distributions. Then the dimensionality reduction method is done by kernel-PCA using the resulting k_{multi} kernel and a latent space z_{kpca} is obtained.
Figure 6.11 shows scatter plots in two dimensions of the obtained latent spaces z_{ssm}, z_{exp} and z_{multi} corresponding to single omics and multi-omic input data by autoencoders and kernel methods.

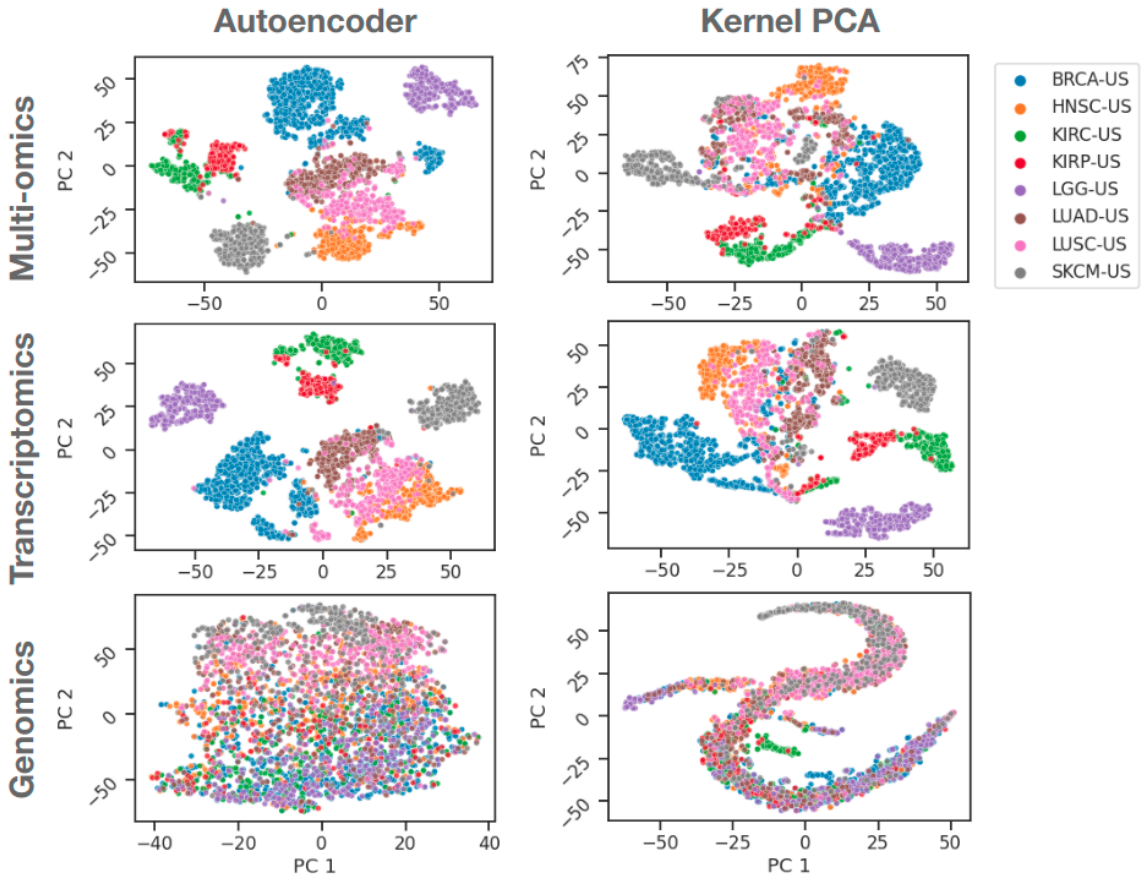

Figure 6.11: Two dimensional visualization via t-SNE embedding of the latent space obtained by each method using single and multi-omics.

It is observed that the genomics modality is noisy since it is difficult to observe separable clusters according to tumor subtypes. On the other side the transcriptomics modality reveals a clear cluster structure between tumor subtypes for both methods, nevertheless the Autoencoder model shows better separability between clusters. Moreover, the resulting z_{multi} obtained by the autoencoder also reveals a better cluster structure in comparison with the one obtained by the kernel PCA method. A possible reason why the autoencoder performs better when compared with the multi-omic kernel approach is the over-parametrized nature of the autoencoder in comparison with the parameter-free kernel model.
As detailed in section 6.3.4 the evaluation of the multi-omic latent space is done by an unsupervised downstream task, particularly a k-means clustering method. It is assumed that the quality of the resulting latent space z_{multi} is determined by how well the downstream task performs. To evaluate the latent space of each single omic and the multi-omic one by the two proposed method the Rand-Index is used.

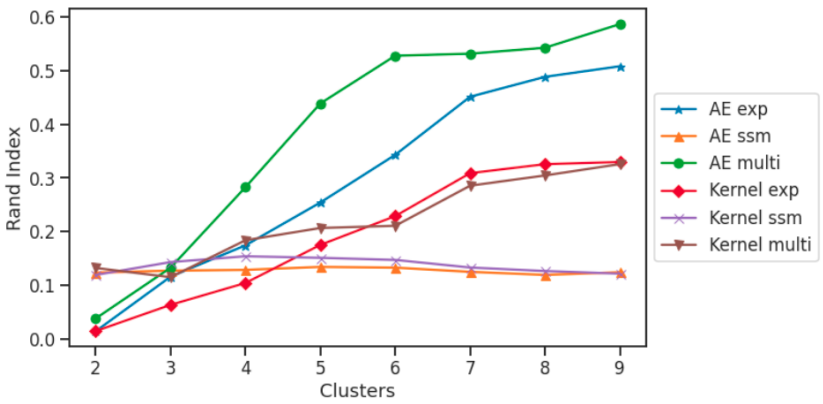

Figure 6.12: Evaluation of single and multi-omic latent space obtained by each integration method.

Figure 6.12 shows Rand Index on the latent space obtained by each method corresponding to each modality for multiple number of resulting clusters. Rand Index is averaged after 10 random iterations of train-test split using the The highest Rand Index is obtained with the Multi-Modal Autoencoder for almost every number of clusters. The results suggest that the latent space z_{multi} learned with the Multi-modal autoencoder improves the accuracy of the clusters according with the tumor subtype labels. Moreover, it is observed that the multi-modal autoencoder learns a latent space that cluster more accurately than the latent spaces obtained by standard autoencoders trained on individual *omics* separately. The kernel combination method reveals a lower Rand Index in the Multi-omic and Transcriptomic domain than the Multimodal Autoencoder. For the genomics domain both methods perform almost the same with the lowest rand index for a number of clusters greater than 4 suggesting that the genomic modality is weakly informative and noisy.

6.6.3 Unsupervised Multi-omic Feature Selection

Once the latent space is learned and the target kernel matrix K_{AE} built then follows the feature selection task via the Multiple Kernel Learning following the LKFS pipeline. To compare the influence of the multi-omic versus single-omic data two latent spaces are learned: one from multi-omic data and another from single-omic data. Then feature selection is performed on gene expression data using as target structure the one learned on each latent space. To evaluate the quality of the selected features the Rand Index is measured after clustering and averaged on 10 train-test splits.

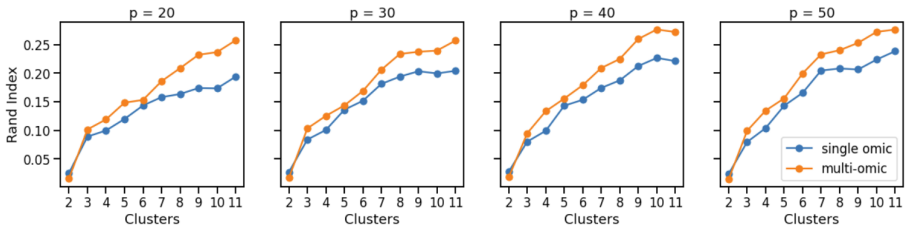

Figure 6.13: Visualization of the distribution of samples on the selected features.

Figure 6.13 compares the Rand Index obtained after clustering using the subset of the selected features. Unsupervised feature selection is done with single omic latent space following the standard procedure of LKFS method and with multi-omic latent space as a new proposed approach. It is observed that the selected features learned from the multi-omic latent space improve the Rand Index therefore the resulting clusters group better the samples of the same tumor subtype toghether without any supervision. The improvement of the Rand Index using the selected features with the multi-omic latent space is consistent for different number of p selected features and shows how the multi-omic data contributes to provide more information about the data structure than the single omic. These results shows that the unsupervised feature selection task on transcriptomics is improved when genomic and transcriptomic data is used to learn the latent space.

6.7 Discussion and Conclusions

This chapter explore and defines the extension of the LKFS method as unsupervised feature selection problem with multimodal latent space in the biomedicine context of multi-omic cancer. Multi-omic data from tumor profiles is more complex than single omic data since the variability comes from different sources simultaneously. To make possible the analysis and interpretation of multi-omic data dimension reduction methods that integrates the different modalities into a single low dimensional representation are needed. To handle this scenario multi-modal dimension reduction methods using Autoencoders and Kernel Methods are presented and described.

The design of the Autoencoder method consists in building an encoder function for each single omic modality, then these functions are merged downstream in the feedforward neural network as a single low dimensional representation $\mathcal{Z}_{\text{multi}}$ known as latent space and detailed in figure 6.2. To reconstruct the input data from the latent space two decoder functions are split. The multi-modal autoencoder tries to reconstruct the multi-omic input data thus the loss function used to learn the network parameters has one term for each omic. Additionally for experimental purposes standard single-modality autoencoders are used on each single omic dataset to compare if the fusion of multiple omic improves the quality of the resulting latent space.

To benchmark the multi-omic integration with autoencoder an alternative to combine and fusion multi-modal data is proposed using a kernel combination approach. It consists in

learning a kernel k_{multi} from the combination of kernels k_{ssm} and k_{exp} corresponding to genomics and transcriptomics modalities respectively. Then by using the k_{multi} in a kernel-PCA algorithm it is possible to learn the low dimensional latent space z_{multi} of the multi-omic domain.

As explained in the materials and methods section the evaluation of the resulting latent space is made by an unsupervised downstream task, more precisely a clustering task by the k-means method. The quality of the resulting latent space is evaluated by how well tumor profiles are clustered after the dimension reduction step according to the tumor subtype labels. With tumor subtypes as ground truth labels a latent space that tends to group togheter tumor profiles of the same subtype is preferred to the ones that tend to distribute different tumor subtypes together.

The resulting multi-omic latent space can be used to extend the LKFS method [129] presented in chapter 05 to multiple omics and improve the feature selection performance. A Multi-omic feature selection task is presented and evaluated. Experimental results are obtained by performing dimensionality reduction with each proposed method using both single omic and multi-omic datasets. To evaluate the latent space the clustering performance is measured. The highest performance is obtained on the latent space learned by the multi-modal autoencoder. The clustering results using the single-omic autoencoder are lower than the one using the multi-modal autoencoder for multiple number of clusters. The results obtained by the kernel combination method are lower than the ones obtained by the autoencoder. Moreover, it is observed that the kernel combination method does not show any improvement between the latent space obtained by single gene expression transcriptomic data and the multi-omic scenario suggesting that the kernel combination method can not capture information from the genomic modality. Future work should consider other types of kernel combination such as $K_{multi}((x,u),(x^{'},u^{'}))$.

The multi-modal autoencoder method outperforms the kernel combination method for the multi-omic data fusion problem. These results suggest that the improvement in the quality of the resulting latent space via the autoencoder in comparison with kernel methods comes with a cost of overparametrization. The kernel integration approach uses just two kernels and that may be a low number of kernels to capture all the underlying structure of the training data that the neural network does. The improvement of the clustering task in the multi-omic latent space of the autoencoder impacts the quality of the feature selection task in comparison with the selected features using a single-omic latent space. Then learning a latent space from multi-omic data improves the corresponding feature selection task using the LKFS. This chapter shows evidence on synthetic and real data about how the latent space can be obtained from multi-omic data and used to improve an unsupervised feature selection task. It is important to remark that in real data the ground truth latent space is not known, further work should consider evaluation tools to determine the quality of the resulting latent space since it impacts directly on the selected features.

In this chapter two approaches have been presented to deal with the multi-modal data fusion problem using synthetic and real pancacer data. An unsupervised multi-omic feature selection method is presented to select features based on a multi-omic latent space. As cancer

multi-omic data repositories keep growing the need to study further and develop multi-modal approaches will be a priority in cancer genomics.

6.7.1 Scientific Production of this chapter

The content of this chapter has been presented in:

- Multimodal Autoencoder has been published as a Journal Paper at BMC Bioinformatics in 2019 titled as *A pan-cancer somatic mutation embedding using autoencoders* [67].

Chapter 7

Conclusions

Throughout this thesis, the different aspects of dimensionality reduction and feature selection applied on tumor profiles characterized by molecular variables such as gene expression level were studied. The dimensionality reduction problem is studied and developed in three approaches.

7.1 Improvement of supervised feature selection and feature extraction via the Kernel Latent Regularization

At the beginning of the thesis in chapter 2 the tumor profiles are considered as been characterized by the expression of thousands of genes. From a supervised machine learning approach, it is possible to predict clinical variables by learning parametric functions that take as input the tumor profile and the expression level of all its genes and generate an output response associated with a clinical variable such as tumor subtype or stage. The supervised learning task in this thesis is classification of tumor profiles into subtypes or estimation of patient survival. The supervised learning problem is applied in a high-dimensional data context caused by the thousands of gene features and it increases the complexity of the problem, therefore the classification performance may be compromised. In order to reduce the dimensionality of the problem and simultaneously achieve interpretability, the feature selection approach is studied. In this work it is developed a method called *Kernel Latent Regularization* to improve the selection of supervised variables. This method consists of selecting a subset of genes that correlate with the clinical variables of interest and simultaneously explain the unsupervised latent structure of the input data. To obtain the latent structure of the data, the *Kernel Principal Component Analysis* method is used and in consequence a low-dimensional unsupervised representation is obtained. Said representation is structured in a pairwise similarity matrix by using Kernel Methods and fused with the labels and clinical variables of the tumors. In this way, a simultaneously supervised and unsupervised hybrid representation is obtained that guides the selection of variables. The selection of variables is carried out by the *Multiple Kernel Learning* method where the linear combination of kernels

associated with each feature results in a final Kernel that helps to find the pairwise similarity between tumor profiles.
The proposed method is evaluated in data from breast, pancreas and lung cancer patients. The results obtained show that the Kernel Latent Regularization improves the classification task after the selection of variables when compared to a purely supervised selection case. In addition, the proposed method is compared with other supervised feature selection methods. The results show that thanks to the proposed method, the classification task is improved by more than 8% in the three types of tumors evaluated on independent test data.
Additionally, in chapter 4, the Kernel Latent Regularization approach is extended to the Sufficient Dimension Reduction methods, a popularly known method for supervised dimensionality reduction. The proposed method is studied in Lung and Breast cancer data showing an increase of up to 4% in the prediction of the patient's survival compared to the purely supervised sufficient reduction.
In this way, it is shown that the role of Kernel Latent Regularization can be used to improve both supervised feature selection and feature extraction tasks by incorporating the latent unsupervised structure of the training data, proposing a new paradigm of supervised and unsupervised learning simultaneously.

7.2 The role of the latent space in the unsupervised feature selection task

Chapter 5 of the thesis studies the case of feature selection for unsupervised learning tasks. The objective is to select a subset of the initial genes that allows tumors to be grouped by similarity in lower dimensionality, assuming that this reduction of dimensions improves the grouping task. To do this, the *Latent Kernel Feature Selection* method is proposed. It first learns a low-dimensional unsupervised latent space from the input data using an unsupervised neural network called *Autocoder*. It is assumed that in the latent space the tumors can be projected in a smaller dimension retaining information about the data distribution. In the latent space obtained, a pairwise similarity matrix between tumor samples is constructed with a Gaussian Kernel function. Then, using the *Multiple kernel Learning* method, genes are selected by a linear combination that best align to the similarity matrix obtained in the previous step. The selected genes are used to characterize the tumor profiles in a considerably smaller dimension where they are grouped according to their similarity.
Additionally, a second unsupervised feature selection method is proposed called *Maximum Mean Discrepancy Latent space feature selection*. This approach uses two *Autocoder* networks to learn a latent space and to select genes respectively. The Maximum Mean Discrepancy criteria makes it possible to bring the latent space distributions of both models similar so that the selected variables approximate the latent structure of the training data.
The two proposed methods are evaluated in data on Brain, Lung and Kidney cancer, in addition, these methods are compared against two other methods of the state of the art in unsupervised feature selection. The experimental results show how the *Kernel Latent Feature*

Selection method selects variables that improve the clustering task more than any of the other methods studied, in some cases with an increase in clustering quality measured by the Rand Index of up to 50%. On the other hand, the method *Maximum Mean Discrepancy Latent Space Feature Selection* has a grouping quality with the selected variables similar to the reference methods. It is important to emphasize that these two methods are designed with a latent space learning stage that guides the selection of variables while the reference methods used in the empirical comparison do not. That is why, through the proposed methods, the role played by the low-dimensional latent space in the selection of variables is significant in improving performance.

7.3 Latent spaces from the multi-omic fusion of tumor profiles.

In chapter 6 of the thesis the feature selection task on a single omic is studied as in chapter 5 with the difference that the dimension reduction from tumor profiles to obtain a latent space is done with multi-modal data sources such as genomics (DNA somatic mutations) and Transcriptomics (gene RNA expression) simultaneously. This implies that each tumor profile is characterized by two data sources simultaneously. The challenge is to be able to achieve a low-dimensional latent representation of these tumors by merging the two available data sources. In the previous chapters cases, the latent spaces are obtained from a single data source which is transcriptomics. The hypothesis proposed in this chapter is that the latent space of tumor profiles can be improved if it is obtained from multiple data sources such as genomics and transcriptomics. To learn a latent space, two approaches are studied: *Multi-omic Fusion with Kernel methods* and *Multi-omic Fusion with multi-modal Autocoders*. The first method creates a pairwise similarity matrix using a kernel function on each available *omic*. Then both matrices are linearly combined to obtain a new matrix that considers the similarity between tumors with the two data sources simultaneously. The resulting matrix is used to perform a dimensionality reduction using the *Kernel Principal Component Analysis* method and obtain a latent space.
The second method is based on Multimodal Autocoders. This method consists of two Autocoders, one for each data source, where each one learns a latent space corresponding to the data source used. Then both models are fully connected to generate a single latent space that contemplates the variability of both data sources.
To evaluate the latent spaces learned from multi-modal data sources by both methods, the tumors are grouped and the quality of the obtained clusters is analyzed based on the clinical attributes, as was done in chapter 5. On this occasion, an experiment is carried out using the following types of tumor: Breast, Neck, Renal, Brain, Lung and Skin. In addition, the quality of the clustering is compared using the separate RNA and DNA data sources in order to observe the effect that learning from multiple sources generates on the quality of the clustering. The results obtained show that the latent space learned from DNA and RNA through Autocoders generates the best clustering quality with up to 20% improvement

over learning using only RNA or DNA omics. In addition, learning a latent space by the multi-modal Autocoder method improves the clustering quality by up to 80% compared to the integration with Kernel methods. The results obtained show that learning a low-dimensional latent space from multiple multi-modal data sources improves the quality of tumor clustering. Clustering enables the discovery of new tumor subtypes as well as finding tumor communities that share clinical attributes.
Finally with the learned multimodal latent space a target kernel matrix is built to guide a feature selection step on transcriptomic data following the Latent Kernel Feature Selection pipeline of chapter 5. Since the multi-modal latent space used as target distribution improves the clustering performance then the selected features in transcriptomics improves the clustering as well.

7.4 Future work

In this thesis four methods of dimensionality reduction for feature selection and feature extraction have been developed to improve both supervised and unsupervised machine learning tasks in biomedical tumor profiles of cancer patients. In all the proposed methods the latent space extracted from the input data improves the learning task for supervised and unsupervised cases. In the synthetic data examples the ground truth distribution is known and it is possible to determine if the resulting latent space is the correct one. Nevertheless in real data scenarios it is not possible to determine the structure of the ground truth latent space. For this reason in real cases if the latent space is not well determined it may limit the quality of the feature selection and consequently the supervised or unsupervised learning task. In this thesis the quality of the latent space is done by a downstream task like classification or clustering. Future work should consider ways to assess the quality of the latent space in real data and to determine how useful or limited is the target representation used for the feature selection task.

Appendix A

Extended abstract in Spanish

A.1 Introduccion

El ritmo creciente de generación de datos de perfiles tumorales durante la última década ha permitido el desarrollo de algoritmos de aprendizaje estadístico para explorar y analizar el panorama de los tipos y subtipos de tumores y la supervivencia de los pacientes desde un punto de vista biomolecular. Los datos tumorales suelen describirse mediante características transcriptómicas y el nivel de expresión de un determinado gen-transcrito en la célula tumoral, por lo que estas características pueden utilizarse para aprender reglas estadísticas que mejoren la comprensión del estado y el tipo de una célula cancerosa.
Sin embargo, los datos transcriptómicos de los tumores son altamente dimensionales y cada tumor puede ser descrito por miles de características genéticas, lo que dificulta la tarea de aprendizaje automático y la comprensión de los mecanismos biológicos subyacentes. Esta tesis estudia cómo reducir la dimensionalidad y ganar interpretabilidad sobre qué genes codifican la señal de la distribución de datos proponiendo métodos de reducción de la dimensión basados en pipelines de selección y extracción de características. Los métodos propuestos se basan en Modelos de Variables Latentes y Métodos Kernel con la idea de explorar la conexión entre las funciones de similitud por pares de las muestras tumorales y los espacios latentes de baja dimensión que capturan la estructura interna de los datos de entrenamiento. Los métodos propuestos han mostrado mejoras en las tareas de selección de características supervisadas y no supervisadas en comparación con los métodos de referencia para clasificar y aprender subgrupos de tumores respectivamente. Además, los métodos propuestos desarrollados en esta tesis han demostrado su adaptabilidad para tratar no sólo con datos de entrada mono-ómicos sino también multi-ómicos.

A.2 Capitulo 01

El primer capítulo de esta tesis, titulado *Reducción de la dimensionalidad en perfiles tumorales biomédicos: un enfoque de aprendizaje automático*, introduce los retos y oportunidades que la

comunidad de aprendizaje automático y reconocimiento de patrones tiene para la exploración de la enfermedad del cáncer en un contexto biomédico. Estas oportunidades se ven impulsadas principalmente por el continuo crecimiento y disponibilidad de grandes datos biomédicos y moleculares de perfiles tumorales como ADN, ARN, Proteínas y Metabolitos [1]. Además, desde el punto de vista matemático y estadístico, los métodos de aprendizaje automático han ido mejorando sus capacidades para procesar datos complejos y de alta dimensión durante las últimas décadas. Finalmente, la mejora de los procesadores computacionales permite la implementación de los métodos de aprendizaje en grandes conjuntos de datos de forma eficiente [2].

A.2.1 Enfermedad del Cancer

El cáncer es un grupo de enfermedades genéticas que pueden desarrollarse en cualquier lugar del cuerpo humano [3]. Se origina cuando células viejas o dañadas que deberían morir, en cambio, sobreviven y crecen sin control alterando la función normal de una célula. Entonces estas células comienzan a dividirse sin parar formando un conjunto de células conocido como tumor. La enfermedad del cáncer se produce como resultado de los cambios y mutaciones producidos dentro de la secuencia de ADN en el genoma de las células cancerosas [4]. El peor escenario se denomina *metástasis* y es cuando el crecimiento de las células tumorales de un determinado órgano continúa y en consecuencia obliga a estas a extenderse a diferentes sitios del cuerpo haciendo que múltiples órganos fallen y finalmente causen la muerte.
La Organización Mundial de la Salud (OMS) estima que 9,6 millones de personas murieron, es decir, una de cada seis muertes en 2018, a causa del cáncer, lo que la sitúa como la segunda causa de muerte a nivel mundial.

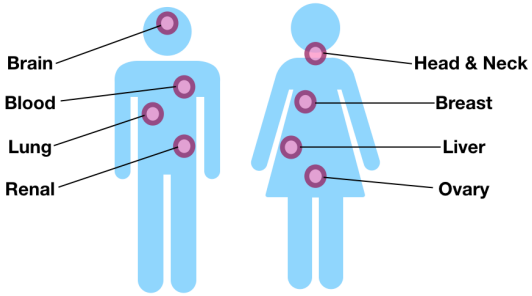

Figure A.1: Primary sites of tumors

A pesar de que cada tumor puede caracterizarse por el mismo conjunto de características biológicas como el ADN, el ARN, las proteínas o los metabolitos, la señal codificada dentro de estos diferirá significativamente dependiendo del sitio primario del cuerpo donde se encuentra el tumor, la etapa del tumor y el subtipo del mismo. Denominamos a esta señal codificada como *firma genética* en referencia a lo activas que son las diferentes regiones y partes del genoma humano dentro de una célula tumoral específica. Por lo tanto, si se miden d características

biológicas, éstas pueden modelarse como variables aleatorias $x_1, x_2, ..., x_i, ..., x_d$ que toman diferentes valores a través de diferentes muestras tumorales regidas por una distribución de probabilidad $p_1(x_1), p_2(x_2), ..., p_i(x_i), ..., p_d(x_d)$. Por ejemplo, un tumor localizado en el Cerebro presentará una firma genética diferente a la de un tumor localizado en el Hígado. Incluso cuando los tumores pertenecen al mismo tipo y sitio primario como el Riñón (Renal) estos pueden ser categorizados en diferentes subtipos como el Carcinoma de Células Claras (KIRC) o el Carcinoma Papilar (KIRP) donde cada uno presenta una firma genética diferente asociada a cada subtipo. Además, la firma genética de un tumor del mismo tipo y subtipo puede diferir en función del *promedio* del paciente, siendo aquellos de baja o alta supervivencia. También es posible que la señal codificada de una determinada firma genética esté localizada en sólo una fracción del genoma humano. Esta situación revela la necesidad de buscar la región o zona del genoma responsable de la codificación de la señal para entender qué mecanismos biológicos están implicados.
Por las razones descritas anteriormente, el procesamiento e interpretación de las firmas genéticas de los tumores es una tarea crucial para comprender la complejidad de la enfermedad del cáncer e identificar con precisión un determinado tumor por tipo, subtipo, estadio o pronóstico. Es importante destacar que, a diferencia de la prevención del cáncer antes de que aparezca como enfermedad, esta tesis se centra en caracterizar con precisión un tumor una vez que ya existe en el paciente. El hecho de poder determinar el tipo, subtipo, estadio o pronóstico asociado de un tumor a partir de una firma genética permite determinar qué tratamiento debe aplicarse a cada paciente en un contexto de medicina personalizada. Además, es posible descubrir bio-marcadores que faciliten la interpretación de los datos genómicos del tumor para mejorar la toma de decisiones en el diagnóstico del cáncer.

A.2.2 Datos tumorales omicos

Cada tumor se caracteriza por diferentes grupos de rasgos moleculares. Cada uno de estos grupos de rasgos está compuesto por el mismo tipo de macromolécula y se denominan *omics*. Cuando un tumor es descrito por diferentes ómicas simultáneamente entonces el contexto se denomina *multi-ómica* [5]. El contexto multiómico descrito se enmarca dentro del concepto de datos multimodales, una situación comúnmente utilizada cuando las características se obtienen de diferentes fuentes y entornos como las imágenes y el lenguaje natural por ejemplo [6]. El contexto multimodal del cáncer implica grupos de características moleculares como el ADN, el ARN, las proteínas o los metabolitos, por lo que cada tumor puede ser descrito por múltiples ómicas de datos biomoleculares.
Además, cada tumor tiene un fenotipo, que es un término asociado a las características y rasgos observables. El fenotipo de un tumor está causado y determinado tanto por las características ómicas como por el entorno. Dado que la enfermedad del cáncer tiene un fuerte componente genético y multiómico, en esta tesis el enfoque y el estudio se centran en las características ómicas del tumor y en cómo estas características moleculares pueden determinar un fenotipo determinado. Los fenotipos tumorales se pueden clasificar por el sitio primario donde se ha desarrollado el tumor y por el tipo histológico [8] que se asocia al tipo

de tejido del tumor. En general, el lugar primario define el tipo de tumor, mientras que el tejido o la histología definen el subtipo de tumor. El tipo de tumor y el subtipo definen el fenotipo que corresponde a un diagnóstico clínico realizado por los médicos clínicos. Además, los datos del tumor están etiquetados por el fenotipo pero también por la información clínica como el estadio del tumor, la supervivencia del paciente, la edad y el sexo del mismo.
Cada ómica contiene potenciales biomarcadores que podrían ayudar a detectar un resultado tumoral como el subtipo de tumor, el estadio o la supervivencia del paciente. Por esta razón, el descubrimiento de biomarcadores es una tarea importante para seleccionar características biomédicas importantes. Además, el descubrimiento de biomarcadores permite a los médicos centrarse sólo en un conjunto reducido y pequeño de características en lugar de medir el genoma completo de un paciente.

A.2.3 Datos tumorales para tareas de aprendizaje automático

En las secciones anteriores se han presentado y descrito los datos ómicos moleculares y los datos clínicos. En esta sección, el objetivo es explicar cómo se pueden utilizar los modelos de aprendizaje automático en este contexto para estimar los resultados clínicos a partir de los datos ómicos. Sea un espacio de entrada $\mathcal{X}$ un espacio d-dimensional tal que $\mathcal{X} \subseteq \mathbb{R}^d$. Entonces un conjunto S_u de n muestras se define en $\mathcal{X}$ como

$$S_u = \{x_1, x_2, ..., x_n\}$$

donde cada muestra x_i es un vector d-dimensional. En esta tesis, cada vector de muestras se asocia a un perfil tumoral caracterizado por un dato ómico específico. Cada conjunto de datos ómicos está compuesto por n perfiles tumorales caracterizados por d rasgos genéticos. Si la información clínica está disponible, entonces un vector de etiquetas y puede definirse como $y \subseteq \mathbb{R}$ para variables clínicas continuas como los días de supervivencia, $y = \{-1, 1\}$ para etiquetas categóricas binarias como el subtipo de tumor o $y = \{1, 2, ..., T\}$ para etiquetas categóricas múltiples como el sitio primario o el estadio del tumor, donde T es el número de etiquetas categóricas clínicas. Entonces, el conjunto de muestras S_u puede redefinirse como un conjunto S_s de pares de entrada-salida de muestras tumorales y etiquetas clínicas

$$S_s = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$$

El conjunto S_u corresponde a un conjunto de muestras no supervisado ya que no hay ninguna etiqueta clínica emparejada con las muestras mientras que el conjunto S_s se considera supervisado ya que cada muestra está emparejada con una etiqueta clínica. En esta tesis propongo tanto los escenarios en los que se dispone de etiquetas clínicas tumorales como aquellos en los que no se dispone de etiquetas, denominados problemas supervisados y no supervisados respectivamente [18]. La figura 2 muestra una viñeta de ejemplo de un conjunto de datos de tumores compuesto por mutaciones somáticas y expresión génica etiquetados con información clínica de los pacientes.

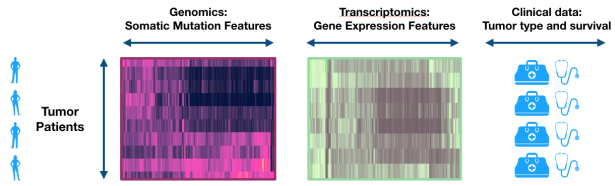

Figure A.2: Diagrama de datos clínicos y multiómicos

En el caso de un enfoque supervisado el objetivo es aprender del conjunto S_s una función $f(x)$ que mapea las muestras de tumor x_i a una etiqueta de tumor y_i como

$$f(x) \sim y \tag{A.1}$$

minimizando una función de pérdida $L(y, f(x))$ que mide la discrepancia entre la salida de $f(x)$ y las verdaderas etiquetas y. Cuando las etiquetas son continuas y $NsubseteqR$ el problema se define como regresión. Cuando las etiquetas son categóricas o discretas $y = \{-1, 1\}$ el problema se define como clasificación. En los casos de contextos no supervisados en los que las etiquetas de los tumores y no están disponibles, el tipo de problemas a resolver está relacionado con el aprendizaje de las similitudes entre las muestras de tumores de S_u y la búsqueda de comunidades o subgrupos de tumores que son altamente similares entre sí. Los grupos resultantes se denominan *clusters* y se espera una correlación entre éstos y las etiquetas clínicas. La figura 3 muestra un ejemplo de problemas de aprendizaje supervisado y no supervisado.

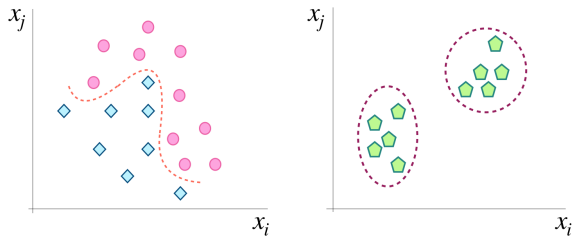

Figure A.3: Figura izquierda: un problema de clasificación supervisado. Figura derecha: un problema de agrupamiento no supervisado.

Además, dado que cada perfil tumoral está caracterizado por más de 20.000 genes codificantes de proteínas, el espacio donde se encuentran las muestras es altamente dimensional. Por lo tanto, una importante tarea de aprendizaje automático que se estudiará en esta tesis es la *reducción de la dimensionalidad*. El objetivo de la reducción de la dimensionalidad es mapear y caracterizar los datos de entrada en una dimensión más baja. Para un algoritmo de aprendizaje automático, los espacios de alta dimensión y los tamaños de muestra bajos aumentan la complejidad del problema, por lo que el aprendizaje de los datos en este

contexto es menos eficiente [19]. Los conjuntos de datos ómicos de tumores utilizados en esta tesis tienen un número de características que supera con creces el número de muestras. Esta situación se conoce como *la maldición de la dimensionalidad* y no es deseable para el aprendizaje estadístico. Además, se asume que el espacio de entrada de alta dimensionalidad contiene varias características ruidosas que pueden ser descartadas por lo que se asume que la dimensionalidad intrínseca de los datos es menor que la de entrada [20].
Por esta razón, el aprendizaje de representaciones de baja dimensión de los datos de entrada y la reducción de la dimensionalidad es una tarea importante que se estudia en esta tesis y hay dos hipótesis importantes a tener en cuenta: en primer lugar, no es necesario aprender del conjunto completo de características biológicas y sólo es necesario un subconjunto reducido de genes para realizar la tarea de aprendizaje automático supervisado o no supervisado. En segundo lugar, existen representaciones latentes más simples y de baja dimensión de los perfiles tumorales que ayudan a entender la distribución y estructura subyacente de los datos.

A.2.4 Desafíos a abordar en esta tesis

Los datos ómicos y clínicos de los tumores presentados invitan a estudiar y desarrollar modelos de aprendizaje automático para evaluar diferentes problemas. La información clínica como el subtipo de tumor, el estadio del tumor o los días de supervivencia pueden estimarse a partir de los datos ómicos en un enfoque supervisado. Además, también se necesitan enfoques no supervisados, ya que no siempre se dispone de las etiquetas clínicas, por lo que es posible descubrir grupos de tumores que presenten una alta similitud entre los perfiles tumorales.
Tanto para los enfoques supervisados como para los no supervisados, los datos ómicos utilizados son de alta dimensión y el número de muestras tumorales es bajo. Este escenario de alta dimensión dificulta la realización de tareas supervisadas o no supervisadas. Hay características genéticas más informativas que otras tanto en los conjuntos de datos de mutaciones somáticas como de expresión génica. Por esta razón, es necesario desarrollar métodos robustos de reducción de la dimensionalidad para extraer información útil de los datos ómicos. La reducción de la dimensionalidad puede realizarse seleccionando un subconjunto de genes p donde $p << d$ o aprendiendo una nueva representación de baja dimensionalidad combinando las características de entrada y filtrando el ruido de entrada.
Esta tesis utilizará los datos de expresión génica sólo como un escenario ómico único mientras que el último capítulo de la tesis estudia la fusión de datos multiómicos que y cómo combinar múltiples ómicas como la mutación somática y la expresión génica.

A.3 Métodos de kernel y selección de características supervisadas

El panorama de los perfiles tumorales está compuesto y categorizado por una amplia gama de clases clínicas diferentes, como los tipos de tumores, los subtipos y la supervivencia de los pacientes, lo que hace que la clasificación de los perfiles tumorales sea una tarea clave

para el aprendizaje automático en la informática del cáncer. La clasificación de tumores implica encontrar una función de decisión matemática utilizada como límite de decisión entre las clases de tumores en un espacio de alta dimensión $\mathcal{X}$ definido por las características de expresión genética de entrada que se consideran como variables aleatorias. La idea central de la clasificación es aprender la función de decisión de forma que asigne con precisión etiquetas clínicas a los perfiles tumorales de contexto clínico desconocido. El problema de la clasificación de perfiles tumorales utilizando datos de expresión génica es altamente dimensional, ya que se caracteriza por tener decenas de miles de características de expresión génica, como se introdujo en el capítulo 01. Además, generalmente los conjuntos de datos de cáncer tienen un tamaño de muestra bajo. Como se mencionó en el capítulo anterior, la relación entre el tamaño de la muestra n y la dimensión del espacio de características d de estos conjuntos de datos es $\frac{n}{d} << 1$, lo que caracteriza situaciones en las que la maldición de la dimensionalidad es un problema de primera importancia. Para abordar un problema de clasificación de tumores en un contexto de alta dimensionalidad, la reducción de la dimensionalidad de los datos es una tarea importante que debe realizarse. La selección de características es una familia de métodos de reducción de la dimensionalidad diseñados para descartar las características ruidosas y no útiles y retener sólo un subconjunto de las características de entrada útiles para realizar una tarea de aprendizaje automático supervisado o no supervisado [22]. Se supone que las características seleccionadas contienen una señal relevante para una tarea de aprendizaje supervisado como la clasificación, mientras que las descartadas contienen ruido o señal redundante. Otro objetivo importante de los métodos de selección de características es guiar el proceso de descubrimiento de biomarcadores [**?**] [24]. Dado que los tumores se caracterizan por decenas de miles de características, tomar medidas de todas las características biológicas es caro e ineficiente. Por otra parte, la selección de un subconjunto de características genéticas puede mejorar una tarea de aprendizaje supervisado como la clasificación de tumores o la regresión de la supervivencia y también mejora una interpretación biomédica, ya que las características seleccionadas proporcionan una manera de entender el sistema biológico subyacente. La selección de características se ha aplicado ampliamente en la bioinformática y la genómica del cáncer utilizando datos de expresión génica en problemas de aprendizaje supervisado [25] [26] [27]. La hipótesis sostenida en este trabajo plantea que es necesario un subconjunto reducido de genes para realizar y mejorar significativamente la tarea de aprendizaje supervisado, en particular en este trabajo la tarea supervisada es la clasificación de tumores.

La tarea de clasificación en la genómica del cáncer toma como entrada un perfil tumoral de expresión génica dado y le asigna una etiqueta de fenotipo mediante la implementación de algoritmos que toman decisiones en función de la optimización de una función de pérdida. Estos algoritmos aprenden reglas estadísticas a partir de datos etiquetados para construir una función de decisión.

Los métodos de aprendizaje supervisado más populares, como las máquinas de vectores de apoyo y el perceptrón, o los no supervisados, como el análisis de componentes principales, utilizan un producto interno $\langle x_i, x_j \rangle$ entre las muestras de datos durante el proceso de aprendizaje, que puede interpretarse como una medida de similitud entre la muestra x_i y

la muestra x_j [28]. A pesar de que esta medida permite el uso de una familia de funciones lineales, éstas pueden ser restringidas y limitadas en aplicaciones del mundo real como los datos tumorales ómicos de alta dimensión. Para ampliar la capacidad de estos métodos, la introducción de métodos kernel permite utilizar este marco lineal para tratar con funciones no lineales más versátiles. Su estudio es el núcleo de esta tesis [18]. Los kernels han sido ampliamente utilizados en biología computacional para tareas supervisadas como la clasificación o la regresión y para tareas no supervisadas como el clustering, así como para resolver problemas de reducción de dimensionalidad [29] [30]. Por lo tanto, cabe esperar que los métodos Kernel también puedan mejorar la tarea de aprendizaje en datos ómicos.
En este contexto surge un conjunto de preguntas: dado un perfil de expresión génica, ¿cómo podemos clasificar las muestras tumorales en función de los atributos clínicos? Si los datos son de alta dimensión, ¿cómo se pueden reducir a un espacio de baja dimensión fácil de procesar? Si los métodos lineales pueden ser limitados en datos tumorales de alta dimensión ¿cómo pueden utilizarse los no lineales para mejorar la tarea de aprendizaje? En este capítulo se utiliza el aprendizaje supervisado en la clasificación de tumores dado un perfil de expresión génica. La selección de características se utiliza para seleccionar un subconjunto de genes que mejore el rendimiento de la clasificación y que ayude a interpretar la firma biológica implicada en la tarea de clasificación. Los métodos Kernel se utilizan para realizar y mejorar tanto la clasificación de tumores como los métodos de selección de características y son la base de un novedoso enfoque propuesto para las tareas de selección de características en los datos de expresión génica.
Este capítulo presenta en primer lugar, en el siguiente orden, los algoritmos de clasificación, los métodos de kernel y los modelos de selección de características. La sección de clasificación define los clasificadores lineales, el método de clasificación de la máquina de vectores de soporte y las métricas de evaluación en la clasificación. La sección de métodos de kernel presenta la métrica de alineación del kernel, el aprendizaje de kernel múltiple, el truco del kernel y cómo los kernels pueden utilizarse para tratar problemas de clasificación en los que las clases no son linealmente separables. A continuación, la siguiente sección trata del enfoque de selección de características, se presentan los métodos de referencia y se definen las métricas de rendimiento. Finalmente, se propone un nuevo enfoque denominado *Kernel Latent Regularization Feature Selection*(KLR-FS) [31] y se detallan los resultados experimentales.

A.3.1 Método propuesto: Selección de características por regularización latente del núcleo

La mayoría de las veces, las características importantes se seleccionan utilizando una función objetivo supervisada [57]. En algunos casos, como la falta de datos, el objetivo supervisado puede ser demasiado estricto y difícil de cumplir para obtener un modelo que pueda generalizar en nuevos datos no vistos [58]. En este escenario limitado, la idea propuesta se basa en considerar la estructura de los datos como otra fuente de aprendizaje además de las etiquetas de la muestra y y utilizarla como estrategia de mejora para las tareas de selección y clasificación de características. Entonces surge una pregunta importante: ¿es posible mejorar el proceso

de selección de características restringiendo el subespacio de características seleccionado no sólo para inferir las etiquetas objetivo sino también para retener más información sobre la estructura de los datos en el espacio de características inicial?
En este trabajo se propone un método de selección de características basado en el Aprendizaje de Núcleos Múltiples (MKL) [43]. Además, el método propuesto combina MKL y un modelo de extracción de características latentes no lineales para mejorar el proceso de selección de características mediante una combinación de enfoques supervisados y no supervisados respectivamente. Esta combinación de enfoques pretende mejorar la capacidad de generalización en la clasificación de las características seleccionadas. Esta estrategia pretende maximizar la separabilidad entre las clases de tumores considerando simultáneamente la estructura latente de los datos de entrenamiento. El método de selección propuesto realiza lo que denominamos una *regularización latente* utilizando simultáneamente las etiquetas de los datos y las variables latentes no supervisadas. Para extraer las variables latentes de los datos de entrenamiento se utiliza un modelo de reducción de dimensionalidad no supervisado, más concretamente el kernel-PCA (kPCA) [59]. La idea clave es buscar características que consideren las etiquetas del tumor y preserven la estructura de los datos simultáneamente. El espacio obtenido debería ser más robusto y conducir a una mejor generalización cuando se trata de tareas de clasificación.
El método propuesto está diseñado para tratar problemas de clasificación de tumores donde la dimensionalidad $d > 18.000$ y el tamaño de la muestra es inferior a $n < 200$ perfiles tumorales. Los perfiles tumorales se clasifican según el estadio o el pronóstico. En este escenario, la mayoría de los algoritmos de selección de características pueden tener un rendimiento inferior debido a la falta de muestras de tumores y al conjunto de características de alta dimensionalidad. La regularización latente propuesta funciona como un proceso de relajación de etiquetas que ha demostrado mejorar el rendimiento de la clasificación de tumores en nuevas muestras de prueba no vistas. El método propuesto se denomina *Kernel Latent Regularization Feature Selection* (KLR-FS).

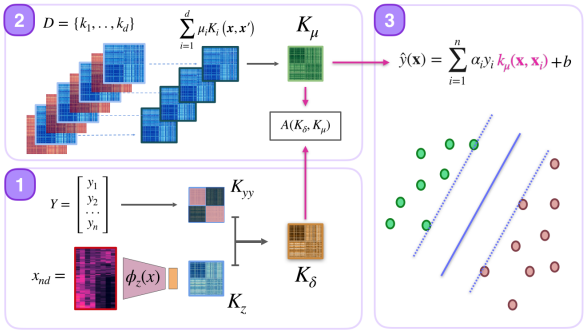

Figure A.4: El proceso KLR-FS. 1) Se construye una matriz K_{yy} utilizando las etiquetas del tumor. Se entrena un modelo kernel-PCA utilizando los datos de entrenamiento y se construye un kernel K_z en el espacio latente Z aprendido del kernel-PCA. La matriz kernel K_{delta} se obtiene de la mezcla entre las matrices K_{yy} y K_z. 2) A partir de los datos de entrenamiento se construye un conjunto D de d kernels de características. Mediante MKL se selecciona un subconjunto de p kernels de características y se obtiene una matriz de kernel K_{mu} mejorando el alineamiento con el kernel K_{delta}. El vector μ indica las características seleccionadas. 3) La función de kernel k_μ se utiliza en la clasificación de vectores de apoyo.

A.3.2 Resultados y discusion

Este trabajo propone un modelo de selección de características basado en el Aprendizaje Kernel Múltiple acoplado al kernel-PCA con una aplicación a la clasificación de tumores utilizando datos de expresión génica de alta dimensión procedentes de perfiles tumorales. El método propuesto KLR-FS tiene como objetivo seleccionar características considerando no sólo las etiquetas de la muestra y sino también las variables latentes z de los datos de entrenamiento presentados en este trabajo como un *regularización latente* ya que se utiliza para mejorar la capacidad de generalización en la selección de características para la clasificación. La idea se basa en considerar no sólo un kernel objetivo supervisado puro k_{yy} sino también un kernel k_z construido a partir de las variables latentes z obtenidas de una reducción de dimensionalidad no lineal ϕ_z. La regularización latente proviene de un kernel construido sobre el espacio latente $\mathcal{Z}$ generado por el kernel-PCA. Esto permite explorar la relajación de las etiquetas de la muestra mediante una combinación lineal de un kernel supervisado y otro no supervisado. Los experimentos muestran que el mayor rendimiento de clasificación y la menor tasa de redundancia en la selección de características para el KLR-FS se obtiene cuando la información latente se mezcla con la información dirigida a la etiqueta. Estos resultados muestran que la selección de variables contemplando tanto la estructura del espacio latente $\mathcal{Z}$ de los datos como las etiquetas del problema supervisado puede mejorar la tarea de clasificación. Además, el aprendizaje de un espacio latente parcial funciona como un término de regularización, ya que limita el espacio de soluciones al introducir la necesidad de capturar también la estructura general de los datos. Al hacerlo, puede mejorar la capacidad de generalización en nuevas muestras de prueba no vistas. Además, existe una relación entre la estructura latente de los datos y las etiquetas de las muestras y y sólo el aprendizaje a

partir de las etiquetas con un criterio supervisado clásico puede dar lugar a un sobreajuste cuando se entrena en un contexto en el que el número de muestras es muy limitado. Por lo tanto, la selección de características con una mezcla de criterios latentes supervisados y no supervisados puede disminuir el sesgo de un clasificador posterior.
KLR-FS produce un kernel personalizado k_{mu} que se utiliza en la clasificación de vectores de soporte y contribuye a mejorar el rendimiento de la clasificación. El kernel k_{μ} resultante puede utilizarse no sólo para tareas de clasificación, sino también para la representación de perfiles tumorales o la visualización mediante kernel-PCA, ya que se construye a partir de fuentes latentes tanto supervisadas como no supervisadas.
El método propuesto tiene relevancia en problemas en los que la estructura de los datos de entrenamiento se correlaciona parcialmente con las etiquetas de los tumores, por lo que aprender también de la estructura latente de los datos mejora la tarea de aprendizaje supervisado.

A.4 Regularización latente del núcleo en métodos de reducción suficientes para el análisis de supervivencia

La reducción de la dimensión es un área importante de problemas en la estadística contemporánea y el aprendizaje automático. Su objetivo es evitar la maldición de la dimensionalidad en las tareas de aprendizaje descendente. Como se ha explicado en los capítulos anteriores, la reducción de la dimensionalidad desempeña un papel fundamental en el análisis del perfil tumoral. El capítulo 01 detalla cómo un solo tumor puede ser caracterizado por decenas de miles de características de expresión genética y la importancia de la reducción de la dimensionalidad. El capítulo 02 estudia la selección supervisada de características para reducir la dimensionalidad manteniendo la interpretabilidad de la reducción y para realizar una tarea de aprendizaje posterior como la clasificación.
En particular, el capítulo 02 propone el método de selección de características de regularización latente del kernel (KLR-FS) [31]. El término de regularización latente en KLR-FS pretende mezclar matrices de kernel supervisadas y no supervisadas para construir un kernel objetivo híbrido que guíe el proceso de selección de características. Por lo tanto, las características seleccionadas son las que capturan información sobre las etiquetas pero también la estructura latente $\mathcal{Z}$ de los datos de entrenamiento. Esta regularización latente muestra cómo la tarea de aprendizaje posterior, como la clasificación, mejora cuando el proceso de aprendizaje considera simultáneamente las etiquetas del tumor y la estructura latente no supervisada. En este capítulo considero interesante explorar la generalización del concepto de regularización latente más allá del método KLR-FS. Una de las características de la regularización latente kernel reside en el hecho de que puede ser utilizada en cualquier tarea de aprendizaje supervisado que tenga una función kernel objetivo durante el proceso de aprendizaje. Por ello, este capítulo pretende ampliar el concepto de regularización latente a otros métodos de aprendizaje que utilicen Kernels como funciones objetivo.
La reducción de la dimensionalidad es comúnmente realizada por enfoques de aprendizaje

no supervisado y es una tarea que puede ser realizada por PCA [73], kernel PCA [74], Autoencoders [75], t-SNE [76] entre otros. Estos métodos aprenden un espacio latente de baja dimensión $\mathcal{Z}$ y tienen la función objetivo diseñada para optimizar la preservación de la estructura de una distribución de entrada X de interés. Por otro lado, los enfoques de aprendizaje supervisado para la reducción de la dimensionalidad se centran en la preservación de la información acerca de una respuesta objetivo y [**?**]. En particular, en este capítulo se estudia el enfoque de Reducción Suficiente (SDR) [78]. El SDR es la familia de métodos supervisados de reducción de la dimensionalidad utilizados para la clasificación y la regresión. Como enfoque supervisado, SDR considera las etiquetas de muestra y para aprender una proyección $\rho(x)$ de los datos de entrada X $en\mathbb{R}^d$ en un espacio de baja dimensión $\mathcal{S}$ $en\mathbb{R}^p$ donde $p < d$ como [79]

$$\rho(\mathcal{X}) = \mathcal{S} \tag{A.2}$$

Formalmente, el subespacio $\mathcal{S}$ es un *subespacio de reducción de dimensión* en términos de la siguiente independencia condicional

$$Y \perp X \mid \beta X \tag{A.3}$$

donde $\perp$ indica independencia y βX representa la proyección ortogonal de X en $\mathcal{S}$. La familia de métodos SDR puede ser útil para predecir etiquetas tumorales como el pronóstico [80]. Los métodos SDR pueden utilizarse para aprender un subespacio $\mathcal{S}$ que capture suficiente información sobre la relación entre las características de la expresión génica y el resultado del fenotipo, como la supervivencia del paciente [81].
Una característica importante de los métodos SDR es que sólo consideran Y para estimar el subespacio $\mathcal{S}$ y no se utiliza la estructura no supervisada de los datos. Este escenario puede provocar un sobreajuste en un contexto en el que el número de características de expresión génica d es mayor que el número de muestras tumorales n ya que $n < d$. La hipótesis estudiada se basa en la idea de que las tareas de aprendizaje supervisado, como la reducción de la dimensionalidad, pueden mejorarse aprendiendo simultáneamente de las etiquetas de las muestras y también de la estructura latente de los datos de entrada. Este capítulo pretende responder a la pregunta: ¿es posible mejorar una tarea de aprendizaje de regresión unida a reducciones de dimensionalidad suficientes introduciendo la regularización latente? Dado que los enfoques de SDR están diseñados para la clasificación y la regresión, en este capítulo se estudia cómo se puede utilizar SDR para el análisis de supervivencia de pacientes mediante el aprendizaje de un espacio supervisado de baja dimensionalidad. La propuesta de este capítulo es un método que mejora la regresión de supervivencia realizada en el espacio proyectado de SDR $\mathcal{S}$ mediante el uso de un método de Regularización Latente Kernel que aprende tanto de las etiquetas del tumor como de la estructura latente no supervisada de los datos de entrenamiento [31].

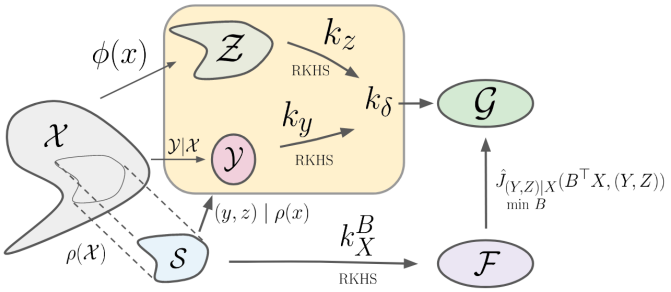

Figure A.5: Método KLR-SDR propuesto. El kernel híbrido k_{delta} mezcla información supervisada y no supervisada de k_y y k_z respectivamente. Entonces, el subespacio de reducción obtenido *mathcalS* explica tanto las etiquetas como la estructura latente de los datos en tránsito.

La figura 5 muestra un diagrama del método propuesto denominado *Kernel Latent Regularization - Sufficient Dimension Reduction* (KLR-SDR). La principal contribución de este capítulo se basa en la extensión del concepto Kernel Latent Regularization propuesto en el capítulo 2 denominado *Kernel Latent Regularization for Feature Selection in Tumor Classification* [31] al método Kernel Sufficient Dimension Reduction (KSDR) [78]. El capítulo de la tesis propone experimentos con datos sintéticos de juguete para la clasificación y regresión de supervivencia para datos reales de expresión génica de tumores. En ambos casos se aprende un espacio latente supervisado de baja dimensión y luego se implementa la tarea de aprendizaje supervisado como la clasificación y la regresión.

A.4.1 Resultados y discusion

Los métodos SDR son métodos de reducción de dimensión adecuados para tareas de aprendizaje supervisado. En particular, el análisis de regresión implica un tipo de modelos capaces de manejar problemas de análisis de supervivencia. Por esta razón, los métodos SDR son adecuados para tratar los datos de supervivencia y de expresión génica de los pacientes con cáncer del Consorcio Internacional del Genoma del Cáncer [90].
La motivación para utilizar la regularización latente del kernel se basa en la idea de tratar con una cantidad limitada de datos (pequeños n), como una forma de preservar la estructura de la clase interna y tener en cuenta la ignorancia sobre el contenido de información del conjunto de datos inicial. Los métodos estándar de SDR no conservan la distribución estructural de los datos tras la reducción de la dimensión y sólo capturan la información discriminante de la distribución condicional $X \mid Y$. Esta situación provoca, en primer lugar, la incapacidad de explorar subclústeres dentro de una clase de tumor después de la reducción SDR y, en segundo lugar, un exceso de ajuste, ya que el aprendizaje de una representación de baja dimensión sólo a partir de las etiquetas en un contexto de bajo tamaño de la muestra puede limitar la representación obtenida a nuevas muestras de prueba no vistas, perdiendo así la capacidad de generalización.

El método estándar de SDR de kernel utiliza un kernel objetivo K_y para guiar la reducción. En este trabajo se propone sustituirlo por el kernel K_{delta} como objetivo para realizar la reducción de dimensionalidad desde un enfoque supervisado a uno no supervisado en el mismo escenario. Además, los experimentos con datos sintéticos para la clasificación binaria revelan cómo se mejora el rendimiento de un clasificador de vectores de soporte cuando se utiliza la regularización latente del kernel en la tarea de reducción de la dimensión.

Los experimentos realizados con datos reales de perfiles de expresión génica de pacientes con cáncer etiquetados con información de supervivencia muestran cómo aprender una representación de baja dimensión $\mathcal{S}$ regularizada por la regularización latente del kernel que finalmente se utiliza para entrenar un modelo de regresión de Cox para la predicción de la supervivencia. La regularización latente del kernel mejora la predicción de la supervivencia en todos los conjuntos de datos en comparación con el caso totalmente supervisado. El método propuesto de regularización latente del kernel para la reducción de la dimensión suficiente (KLR-SDR) con el parámetro de mezcla δ, que alcanza el máximo rendimiento de la regresión, se compara con los métodos kSDR, PCA y kPCA. El método KLR-SDR supera a todos los métodos de referencia mostrando el potencial del módulo de regularización latente. En este trabajo se utiliza el método kernel-PCA para extraer un espacio latente $\mathcal{Z}$ para el término de regularización latente, sin embargo se puede utilizar cualquier otro método no supervisado como t-SNE, Autoencoders o PCA.

Los resultados obtenidos tanto en datos sintéticos como en datos reales muestran que el aprendizaje no sólo de las etiquetas supervisadas sino también de la estructura latente no supervisada mejora la tarea de aprendizaje supervisado posterior. Los resultados sugieren que la regularización latente mejora tanto las tareas de clasificación como las de regresión, por lo que el enfoque de regularización latente de kernel propuesto podría generalizarse a cualquier método de aprendizaje supervisado que utilice kernels para guiar el proceso de aprendizaje.

La regularización de Kernel Latent mejora un método de reducción de dimensión suficiente de Kernel para problemas de regresión de supervivencia de pacientes con cáncer. Los resultados sugieren que el aprendizaje a partir de la estructura latente de los datos de entrenamiento mejora las tareas de aprendizaje supervisado. La motivación de utilizar la estructura interna de los datos durante el entrenamiento se basa en tener en cuenta la ignorancia del contenido de información del conjunto de datos inicial, ya que la baja densidad de datos puede no permitir el descubrimiento de subgrupos de clase internos. Además, el uso de la estructura de datos mezclada con etiquetas mejora la generalización de una tarea supervisada en nuevos datos de prueba no vistos.

Los métodos de kernel han demostrado el potencial de mejora incluso en escenarios complejos cuando el número de características supera con creces el número de muestras disponibles $d > n$ como los datos genómicos del cáncer. En futuros trabajos se considerarán los datos multiómicos de los pacientes.

A.5 Selección de características no supervisada con autocodificadores y métodos kernel

Este capítulo está dedicado a explorar y comprender los métodos de selección de características y de reducción de la dimensionalidad a partir de los datos de expresión génica de los perfiles tumorales con el objetivo de descubrir biomarcadores y subtipos de tumores mediante métodos de aprendizaje automático no supervisado. El capítulo 02 define los problemas de aprendizaje supervisado y los métodos de selección de características para mejorar una tarea de clasificación. Un aspecto principal del capítulo 02 es la disponibilidad de etiquetas tumorales y como el subtipo o la supervivencia durante el entrenamiento del modelo que permite la aplicación de clasificadores como las máquinas de vectores de soporte. Además, en el capítulo 02 las etiquetas de los tumores se utilizan durante el proceso de selección de características, por lo que no sólo se supervisa la clasificación sino también la selección. Sin embargo, las etiquetas de los tumores no siempre están disponibles durante el proceso de aprendizaje y en estos casos se necesita un método no supervisado. El aprendizaje no supervisado es una amplia rama de los métodos de aprendizaje automático que abarca desde la reducción de la dimensionalidad hasta la agrupación. Además, el contexto altamente dimensional de los datos genómicos de los perfiles tumorales requiere reducir la dimensionalidad para mejorar el rendimiento de una tarea de aprendizaje no supervisado posterior. Una de las preguntas que este capítulo trata de responder es: ¿es posible seleccionar genes sin etiquetas clínicas de forma que se facilite el descubrimiento de subtipos de tumores? Este capítulo formaliza y explica cómo seleccionar un subconjunto reducido de genes que mejore el clustering tumoral como forma de guiar el descubrimiento de nuevos subtipos tumorales [96]. La interpretación del sistema biológico se realiza a través del descubrimiento de biomarcadores con métodos de selección de características y el descubrimiento de subtipos a través de una tarea de clustering. Cada perfil tumoral está descrito por más de $d = 10.000$ características de expresión génica en comparación con un tamaño de muestra relativamente bajo n que define un espacio de entrada de alta dimensión y una mayor complejidad en el análisis de datos. Además, los tipos de tumores presentan una heterogeneidad interna que puede subdividirse en subgrupos con rasgos clínicos comunes y que se denominan subtipos de tumores, por lo que el descubrimiento de subtipos de tumores es una tarea importante en la genómica del cáncer, ya que permite a los médicos clínicos asignar un diagnóstico y unos tratamientos diferentes incluso dentro del mismo tipo de tumor [97] [98] [**?**]. Dado el alto espacio dimensional de los datos de entrada es necesario reducir la dimensionalidad mientras se preserva la interpretación biológica del sistema como una firma genética reducida. Estas razones motivan este capítulo para proponer dos nuevos métodos de selección de características no supervisadas para la agrupación de tumores.

Como se explica en el capítulo 02, los métodos de selección de características son necesarios en la genómica del cáncer, ya que proporcionan una representación de baja dimensión de los datos de entrada caracterizada por un subconjunto seleccionado de los genes de entrada que proporciona interpretabilidad de los resultados y descarta el resto siguiendo una función objetivo relacionada con la mejora de una tarea de aprendizaje [100]. Además, los genes

seleccionados pueden utilizarse para guiar las estrategias de descubrimiento de biomarcadores [24]. El subespacio resultante obtenido por un método de selección de características es descrito explícitamente por un subconjunto de características biológicas asumiendo que el conjunto de características inicial contiene características ruidosas que pueden ser descartadas. Un subconjunto de características reducido tiene beneficios en la reducción de la complejidad del modelo y en la medición de sólo un conjunto reducido de biomarcadores [26].
En un problema no supervisado, ya que las etiquetas y no están disponibles durante el entrenamiento, se espera que los genes seleccionados puedan revelar la estructura de los datos, lo que se evalúa en una mejora de la agrupación de los perfiles tumorales. Para guiar el proceso de selección de características este trabajo propone un método que primero aprende una representación poco dimensional y denotada de los datos de entrada conocida como espacio latente $\mathcal{Z}$ que es aprendida por una red neuronal no supervisada conocida como Autoencoder. En este capítulo se proponen dos estrategias. La primera es utilizar un modelo de Aprendizaje de Núcleo Múltiple para seleccionar un subconjunto de características genéticas con el objetivo de alinear lo más posible la distribución resultante de las características seleccionadas a una distribución objetivo de las muestras de entrenamiento en la representación aprendida $\mathcal{Z}$. La segunda estrategia en lugar de utilizar un método de Aprendizaje Kernel Múltiple para la tarea de selección de características, utiliza un autoencoder de vainilla como una red de estudiantes para aprender un segundo espacio latente de baja dimensión $\mathcal{Z}_v$ de manera que la discrepancia de la distribución de datos en $\mathcal{Z}_v$ contra la del espacio latente $\mathcal{Z}$ se minimiza mediante el uso de la distancia de máxima discrepancia media (MMD) [101]. Finalmente, en ambos métodos propuestos, al realizar el clustering sobre el subespacio obtenido a partir de las características seleccionadas, se espera observar atributos clínicos significativos asociados a cada cluster y así validar la calidad de las características seleccionadas.
La principal contribución de este capítulo son dos métodos no supervisados capaces de seleccionar genes con relevancia clínica a partir de datos de expresión génica de alta dimensión sin necesidad de tener etiquetas tumorales.

A.5.1 Aprendizaje no supervisado

El objetivo del aprendizaje no supervisado es capturar la estructura de la distribución de los datos basándose en la similitud entre las muestras de datos. La diferencia clave entre el aprendizaje supervisado y el no supervisado es que este último no utiliza ninguna etiqueta de respuesta y durante el entrenamiento de un modelo. En esta línea, la selección de características no supervisada se basa en encontrar un subconjunto de características de entrada que capture la estructura de los datos también sin utilizar ninguna etiqueta de tumor. En el contexto de la genómica del cáncer, se supone que los datos de expresión genética son una variable aleatoria independiente representada por una matriz de entrada $\mathbf{X}$. La matriz de entrada $\mathbf{X}$ representa un conjunto de n vectores $d-$ de entrada x_i, $i = 1...N$ y x_i es la $i^{\underline{a}}$ muestra del conjunto de datos. Cada vector de entrada x_i está asociado a una muestra de tumor y cada dimensión j del vector de entrada x_{ij} es un gen y representa el nivel de expresión correspondiente. Los vectores de entrada son muestras aleatorias de valor real

obtenidas de un espacio $d-$dimensional $\mathcal{X} en \mathbb{R}^d$. Entonces, dado un conjunto S de n muestras

$$S = \{x_1, x_2, ..., x_n\} \tag{A.4}$$

el objetivo del aprendizaje no supervisado es describir la estructura, la distribución de los datos y la similitud entre las muestras del conjunto S.
La motivación de utilizar el aprendizaje no supervisado para la selección de características en los perfiles tumorales se basa en dos supuestos. El primer supuesto establece que dentro de una clase de tipo de tumor conocido pueden existir subgrupos internos y heterogeneidad interna de importancia clínica. Por lo tanto, se asume que un conjunto de datos de tumores puede presentar grupos o clusters de muestras tumorales con alta intra-similitud que no están descritos por una etiqueta clínica, por lo que la búsqueda de estos subgrupos es deseable. La segunda suposición se basa en la idea de que la distribución y la estructura de los datos de entrada tienen un *dimensionalidad intrínseca* p menor que la dimensionalidad inicial de entrada d. Por esta razón, se supone que es posible describir la variabilidad y la estructura de los datos de entrenamiento en una dimensión inferior.
En este trabajo los enfoques de *clustering* [102], *modelos de variables latentes* [60] y *métodos de kernel* [18] se utilizan para resolver los problemas de encontrar subgrupos y reducir la dimensionalidad de los datos de entrada. Estos métodos se combinan para realizar una selección de características no supervisada en los datos de expresión génica de los perfiles tumorales.

A.5.2 Metodo propuesto

Para realizar la selección de características no supervisada en este capítulo se presentan dos métodos. El primero es la selección de características de kernel latente. Se compone de dos modelos: un autocodificador y un método de aprendizaje de núcleos múltiples. El primero, el autoencoder, se construye para aprender un espacio latente de baja dimensión. A continuación, se diseña la tarea de Aprendizaje de Núcleo Múltiple para seleccionar las características que siguen la distribución de datos del espacio latente, como se muestra en la figura A6.

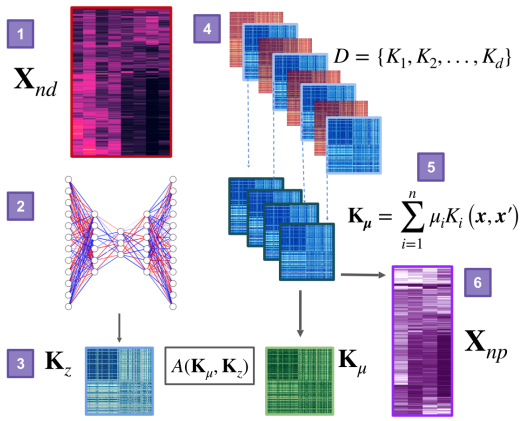

Figure A.6: Tamaño de la matriz de datos $\mathbf{X}$ antes (izquierda) y después (derecha) de la selección de características.d es el número de características de entrada y p es el número de características seleccionadas.

El segundo método propuesto en este capítulo se basa en dos autocodificadores y se denomina Selección de características de máxima discrepancia media latente (LMMD-FS). El primero se encarga de aprender un espacio latente de baja dimensión de los datos de entrada como hace el método LKFS. El segundo autocodificador es un autocodificador de vainilla de una sola capa responsable de seleccionar las características que mejor se aproximan a las características latentes obtenidas a partir del autocodificador. La figura A7 muestra un diagrama del método propuesto.

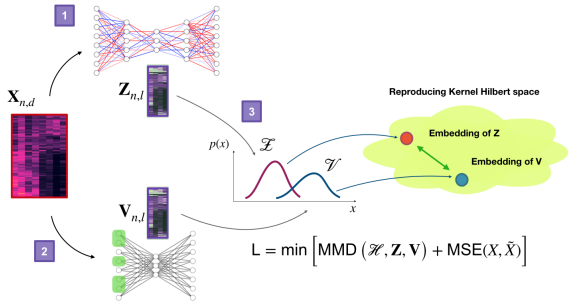

Figure A.7: Tamaño de la matriz de datos $\mathbf{X}$ antes (izquierda) y después (derecha) de la selección de características.d es el número de características de entrada y p es el número de características seleccionadas.

A.5.3 Discusión y conclusión

Este trabajo propone los métodos LKFS y LMMD-FS en los que ambos utilizan un enfoque no supervisado para seleccionar un subconjunto de genes reducido de más de 8.000 genes a menos de 50 en tres tipos diferentes de conjuntos de datos de cáncer con el objetivo de mejorar el rendimiento del clustering. La selección de las características LKFS se realiza mejorando

la alineación entre una matriz kernel K_z obtenida del espacio latente $\mathcal{Z}$ aprendido mediante un autoencoder y por la resultante K_μ tras el aprendizaje de kernel múltiple en kernels de características. La selección del LMMD-FS se realiza minimizando la discrepancia MMD entre el espacio latente de dos autoencodificadores y construyendo un ranking de características basado en la norma L1. Ambos enfoques propuestos se basan en la construcción de una representación objetivo de los datos de entrada con el espacio latente del autocodificador profundo regularizado. Esta representación objetivo tiene un nivel reducido de ruido y es poco dimensional. LKFS y LMMD-FS seleccionan sólo las características que más se alinean mediante la alineación del kernel y reducen la distancia MMD a la representación objetivo, respectivamente. La representación objetivo puede interpretarse como una distribución a priori que guía el proceso de selección.

Para medir la calidad del proceso de selección de características sin supervisión, se calcula la puntuación de redundancia RED en las características seleccionadas. Además, se aplica un método de agrupación de K-means como tarea posterior y el rendimiento de los clústeres se evalúa mediante la homogeneidad de las etiquetas de la verdad básica en los clústeres mediante el índice de Rand.

A partir de los resultados experimentales se observa que LKFS supera a los métodos LMMD-FS y bechmark en el índice Rand y la puntuación RED para casi todos los conjuntos de datos y el número de características seleccionadas. Esto sugiere que LKFS es capaz de seleccionar características de baja redundancia a partir de un espacio de entrada de alta dimensión que contribuye a encontrar clusters bien definidos compuestos principalmente por un subtipo de tumor. Además, el método LMMD-FS tiene un comportamiento similar al de los métodos de referencia en casi todos los conjuntos de datos para cualquier número de clusters, con la excepción del conjunto de datos Brain, donde el índice Rand obtenido es el segundo más alto.

Una de las ventajas tanto de LKFS como de LMMD-FS es la representación latente objetivo aprendida por el autoencoder. Esta representación captura las características más destacadas del conjunto de datos de entrada y luego, mediante MKL o MMD, las características seleccionadas capturarán aproximadamente la misma estructura de datos. Los métodos LKFS y LMMD-FS proporcionan dos salidas. La primera es el subconjunto de características genéticas seleccionadas, que se reduce considerablemente en comparación con el conjunto de características original y ayuda a la interpretación biológica. El segundo es el espacio latente proporcionado por el autoencoder. El espacio latente no sólo sirve como representación del objetivo para el proceso de selección de características, sino también como herramienta para la exploración y el análisis de datos, ya que puede resumir en un espacio de menor dimensión las características más destacadas de los datos originales.

Una de las limitaciones de LKFS y LMMD-FS consiste en entrenar dos modelos, el autocodificador profundo regularizado y el autocodificador MKL o Vanilla, respectivamente. A continuación, los enfoques de selección de características están condicionados por la arquitectura y la calidad del autocodificador profundo regularizado.

A.6 Fusion multi-omica en perfiles tumorales

Los datos moleculares de los perfiles tumorales pueden dividirse en múltiples grupos de variables aleatorias asociadas a las mediciones realizadas en contextos específicos. Cada uno de estos grupos es una fuente específica de variabilidad medida en diferentes contextos moleculares. El escenario multimodal está descrito por fuentes de variabilidad conocidas como *ómicas* donde cada *ómica* está asociada a un tipo particular de características moleculares: genoma, transcriptoma, proteoma y metaboloma [5]. Cada ómica está compuesta por macromoléculas que codifican información biológica sobre la función celular siguiendo el dogma de la biología central y tiene un impacto diferente en la regulación biológica dentro de la célula tumoral.

Para obtener datos de los diferentes entornos ómicos se utilizan diferentes tecnologías. Por ejemplo, la tecnología de secuenciación de nueva generación (NGS) de alto rendimiento mencionada en el capítulo 01 se utiliza para obtener la genómica (mutaciones de ADN) y la transcriptómica (concentración de ARN), mientras que la tecnología de espectrometría de masas [135] se utiliza para obtener datos de la proteómica (concentración de proteínas).

A pesar de que el coste de la secuenciación de las diferentes ómicas está disminuyendo globalmente la complejidad para obtener múltiples ómicas para cada muestra de cáncer sigue siendo una limitación causada por el equipo o la capacidad del laboratorio por esta razón algunos experimentos obtienen datos de una sola ómica, comúnmente *transcriptómica* y una carencia del resto de las ómicas. En esta tesis desde el capítulo 02 hasta el capítulo 04 se ha utilizado la *transcriptómica* o expresión génica ya que es una fuente de información útil para determinar el estado de una determinada célula y es una de las ómicas más utilizadas en el análisis de células individuales y en la genómica del cáncer [136] |cascianelli2020machine [138].

Sin embargo, si un conjunto de perfiles tumorales es descrito por múltiples ómicas en lugar de aprender de una sola, es posible proponer métodos que aprendan simultáneamente de diferentes ómicas para mejorar una tarea de aprendizaje específica. Se supone que cada ómica proporciona información específica sobre la distribución de los datos, por lo que el aprendizaje a partir de múltiples ómicas puede explicar mejor la variedad de datos.

Dado que los datos ómicos individuales son de alta dimensión, los datos multiómicos son incluso de mayor dimensión, este capítulo estudia y analiza si es posible aprender un espacio latente de baja dimensión cuando se utilizan múltiples ómicos como datos de entrada y fusionarlos en una única representación. Para aprender un único espacio latente de baja dimensión $\mathcal{Z}$ a partir de datos multiómicos se puede utilizar el enfoque *Multi-Modal Learning* relacionando la información de múltiples fuentes y como consecuencia mejorar una tarea de aprendizaje descendente en el espacio latente [139] con la suposición de que cada ómico puede explicar alguna estructura particular de los datos por su cuenta. En este trabajo la tarea de aprendizaje descendente es el clustering.

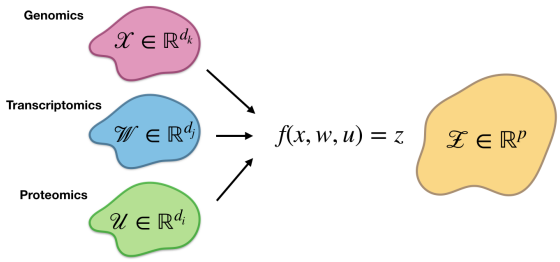

Figure A.8: Esquema del problema de fusión multiómica. A partir de múltiples ómicas $\mathcal{X}$,$\mathcal{W}$,$\mathcal{U}$ se aprende una representación de baja dimensión $\mathcal{Z}$.

El enfoque de aprendizaje multimodal y la fusión de datos multimodales pueden implementarse mediante diferentes estrategias. Se puede hacer mediante el uso de Autoencoders [139][140] o mediante el uso de métodos de kernel, más precisamente la combinación de kernels y kernel-PCA [**?**].
Una vez aprendido el espacio latente a partir de los datos multiómicos de entrada, se puede realizar una tarea de selección de características no supervisada sobre las características transcriptómicas de entrada, tal y como se hizo en el capítulo 03 con el método de selección de características de kernel latente (LKFS) [129] donde el objetivo es guiar mediante un espacio latente la selección de un subconjunto de las características de entrada que mantenga la estructura de datos necesaria para encontrar clusters o comunidades de tumores con atributos clínicos comunes. La principal diferencia de este trabajo con respecto al método LKFS es que el espacio latente aprendido representa una estructura de datos procedente de múltiples ómicas en lugar de una única. Si el espacio latente aprendido a partir de datos multiómicos tiene una mejor estructura que un espacio latente aprendido a partir de datos de una sola ómica basado en el rendimiento de la agrupación, entonces la tarea de selección de características no supervisada de LKFS puede mejorarse si se utilizan múltiples ómicas para el aprendizaje del espacio latente. En este trabajo se utilizan dos ómicas en la sección experimental: Mutaciones somáticas simples (genómica) y RNA-seq (transcriptómica). Para entender el impacto de hacer el aprendizaje del espacio latente y la selección de características no supervisadas, se evalúan dos escenarios: una sola ómica en la que todas las muestras se describen sólo por las características transcriptómicas y una multiómica en la que todas las muestras se caracterizan simultáneamente por la transcriptómica y la genómica.

A.6.1 Fusion de datos multi-omicos

Los métodos de fusión multiómica tienen como objetivo aprender una representación de baja dimensión de las muestras tumorales caracterizadas por múltiples ómicas. La representación de baja dimensión captura la variabilidad de las múltiples ómicas de entrada en una única representación $\mathcal{Z}$. Esto significa que si el método de fusión multiómica está bien implementado, se espera que capture la estructura de los datos procedentes de todas las ómicas de entrada, por lo que la fusión multiómica puede explicar la estructura que una sola ómica no puede por

sí sola.
Como se detalla en la sección de trabajos relacionados, en este capítulo se estudian dos alternativas para la fusión de datos multiómicos. La primera es a través de autocodificadores multimodales en los que los datos multiómicos se utilizan para aprender un único espacio latente de baja dimensión $\mathcal{Z}$ a través de una función codificadora y se reconstruyen desde $\mathcal{Z}$ a $\mathcal{X}$ a través de una función decodificadora. La segunda alternativa utilizada en este trabajo es la combinación de kernels, cada uno de ellos utilizado en cada ómica de entrada y luego fusionado en un único kernel cuya representación correspondiente captura la variabilidad y la estructura de todas las ómicas.
En este trabajo se utilizan ambas estrategias para aprender una representación conjunta a partir de datos de Mutación Somática Simple (Genómica) y de Secuencia de Expresión Génica (Transcriptómica) de perfiles tumorales. Los perfiles tumorales se describen mediante variables aleatorias procedentes de ambas ómicas donde d_x y d_u son las dimensiones que definen los espacios $\mathcal{X} \in \mathbb{R}^{d_x}$ y $\mathcal{U}\mathbb{R}^{d_u}$ respectivamente. Por esta razón, una muestra de tumor puede ser descrita por un solo ómico si sólo se dispone de uno de ellos o por ambos en caso de que se disponga de los dos. Por lo tanto, un conjunto X de muestras caracterizadas por una sola ómica $\mathcal{X}$ se define como

$$X = \{x_1, \ldots, x_n\}$$

o caracterizado sólo por $\mathcal{U}$ definido como

$$U = \{u_1, \ldots, u_n\}$$

y, por otro lado, se define un conjunto de muestras multiómicas cuando se dispone de las dos ómicas como

$$M_{x,u} = \{m(x_1, u_1), \ldots, m(x_n, u_n)]\}$$

donde n corresponde al número de muestras, x las muestras descritas sólo por la primera ómica, u las muestras descritas sólo por la segunda ómica y m las descritas simultáneamente por las dos ómicas. En la sección experimental se evalúan los métodos de reducción de la dimensionalidad utilizados en conjuntos de datos mono-ómicos y multi-ómicos con el fin de analizar cómo el aprendizaje de datos multi-ómicos mejora una tarea posterior como el clustering.
En los siguientes subapartados se explica la metodología utilizada para fusionar los datos multiómicos en una única representación $\mathcal{Z}$ mediante Autoencoders y Kernel Methods.

A.6.2 Método propuesto: Selección de rasgos multiómicos latentes

El capítulo 03 de esta tesis propone el método Latent Kernel Feature Selection (LKFS) diseñado para realizar una selección de características no supervisada en datos ómicos simples. En esta sección primero se describe brevemente el método LKFS y luego se extiende a los datos multiómicos.
El método LKFS está diseñado para aprender primero un espacio latente $z = \phi(x)$ en el que

el mapeo ϕ es un autoencoder y x son los datos de entrenamiento de un solo ítem. En el espacio latente aprendido se aprende un kernel $k_{AE}(z_i, z_j)$ y se utiliza para construir una matriz kernel K_{AE} que detalla las medidas de similitud por pares entre las muestras que se encuentran en $\mathcal{Z}$. A continuación, a partir de un conjunto D de matrices de kernel de características K_i un enfoque de aprendizaje de kernel múltiple realiza una combinación lineal $K_\mu = \sum \mu_i K_i$ con el objetivo de maximizar la alineación $A(K_{AE}, K_\mu)$ donde K_μ es la matriz de kernel resultante de la combinación lineal y μ es el vector solución. Para maximizar la función objetivo $A(K_{AE}, K_\mu)$ el vector solución puede ser escaso, lo que implica que las características seleccionadas son aquellas en las que $\mu_i > 0$.
Para ampliar el LKFS a los datos multiómicos se utiliza un conjunto de datos $M = [m(x_1, u_1), ..., m(x_n, u_n)]$ para entrenar un autocodificador multimodal en lugar del autocodificador de una sola modalidad. Se obtiene un espacio latente $\mathcal{Z}$ y se utiliza como kernel objetivo para guiar la tarea de selección de características con el paso de aprendizaje de kernel múltiple.

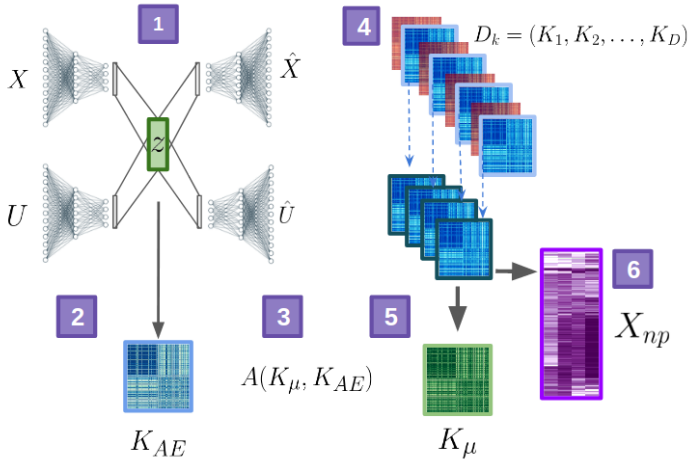

Figure A.9: Propuesta de línea de producción para la selección no supervisada de características pan-ómicas del cáncer. Primero se aprende un espacio latente y se construye una matriz de kernel objetivo K_{AE}. A continuación, se utiliza un conjunto D de núcleos de características para resolver un problema de aprendizaje de núcleos múltiples en el que el núcleo resultante K_μ mejora la alineación $A(K_{AE}, K_\mu)$. Finalmente se obtiene un subconjunto de características p considerando las características donde el vector solución es $\mu_i > 0$.

Las características se siguen seleccionando a partir de una sola ómica, en este caso la transómica. Sin embargo, el espacio latente se aprende a partir de la multiómica. Esto puede provocar que las características seleccionadas de la transcriptómica respondan a la estructura obtenida de la multiómica y que, finalmente, cambien la solución en comparación con la tarea de selección de características de una sola ómica.
La figura 4 muestra el proceso del método de selección de características multiómicas no supervisado propuesto.

A.6.3 Resultados y discusion

Este capítulo explora y define los datos problemáticos multimodales en el contexto biomédico del cáncer multiómico. Los datos multiómicos de los perfiles tumorales son más complejos que los datos ómicos simples, ya que la variabilidad proviene de diferentes fuentes simultáneamente. Para hacer posible el análisis y la interpretación de los datos multiómicos se necesitan métodos de reducción de dimensión que fusionen las diferentes modalidades en una única representación de baja dimensión. Para manejar este escenario se presentan y describen métodos de reducción dimensional multimodal utilizando Autoencoders y Métodos Kernel.
El diseño del método Autoencoder consiste en construir una función codificadora para cada modalidad ómica individual, luego estas funciones se fusionan aguas abajo en la red neuronal feedforward como una única representación de baja dimensión $z_{textmulti}$ conocida como espacio latente y detallada en la figura 2. Para reconstruir los datos de entrada a partir del espacio latente se dividen dos funciones decodificadoras. El autocodificador multimodal intenta reconstruir los datos de entrada multiómicos, por lo que la función de pérdida utilizada para aprender los parámetros de la red tiene un término para cada ómica. Además, con fines experimentales, se utilizan autocodificadores estándar de una sola modalidad en cada conjunto de datos de un solo ómico para comparar si la fusión de múltiples ómicos mejora la calidad del espacio latente resultante.
Otra alternativa propuesta para combinar y fusionar datos multimodales es el método de combinación de kernels. Consiste en aprender un kernel $k_{textmulti}$ a partir de la combinación de kernels $k_{textssm}$ y $k_{textexp}$ correspondientes a las modalidades genómica y transcriptómica respectivamente. A continuación, utilizando el $k_{textmulti}$ en un algoritmo kernel-PCA es posible aprender el espacio latente de baja dimensión $z_{textmulti}$ del dominio multiómico.
Como se explica en la sección de materiales y métodos, la evaluación del espacio latente resultante se realiza mediante una tarea descendente no supervisada, más concretamente una tarea de clustering mediante el método k-means. La calidad del espacio latente resultante se evalúa en función de lo bien que se agrupan los perfiles tumorales después de la etapa de reducción de la dimensión según las etiquetas de los subtipos tumorales. Con los subtipos tumorales como etiquetas de verdad, se prefiere un espacio latente que tiende a agrupar perfiles tumorales del mismo subtipo a los que tienden a distribuir diferentes subtipos tumorales juntos.
El espacio latente multiómico resultante puede utilizarse para extender el método LKFS [129] presentado en el capítulo 03 a múltiples ómicas y mejorar el rendimiento de la selección de características. Se presenta y evalúa una tarea de selección de características multiómicas. Los resultados experimentales se obtienen realizando una reducción de la dimensionalidad con cada método propuesto utilizando conjuntos de datos mono-ómicos y multi-ómicos. Para evaluar el espacio latente se mide el rendimiento del clustering y se observa que el mayor rendimiento se obtiene en el espacio latente aprendido por el autoencoder multimodal. Los resultados obtenidos con el autocodificador monomodal son inferiores a los obtenidos con el autocodificador multimodal entrenado en los conjuntos de datos multimodales. Los resultados obtenidos por el método de combinación de kernels son inferiores a los obtenidos por el autocodificador. Además, se observa que el método de combinación de kernels no muestra

ninguna mejora entre el espacio latente obtenido por los datos transcriptómicos de expresión de un solo gen y el escenario multiómico, lo que sugiere que el método de combinación de kernels no puede capturar la información de la modalidad genómica. El trabajo futuro debería considerar otros tipos de combinación kernel como $K_{multi}((x,u),(x^{'},u^{'}))$.
Dado que el método de autoencoder multimodal supera al método de combinación de núcleos para el problema de fusión de datos multiómicos, estos resultados sugieren que el enfoque de autoencoder mejora la calidad del espacio latente resultante en comparación con los métodos de núcleos con un coste de sobreparametrización. La mejora de la agrupación en el espacio latente repercute en la calidad de la tarea de selección de características, mostrando que las características seleccionadas a través del espacio latente multi-ómico superan a las seleccionadas a través de un espacio latente mono-ómico.
En este capítulo se han presentado dos enfoques para tratar el problema de la fusión de datos multimodales utilizando datos sintéticos y reales de pancacer y se presenta un método de selección de características multiómicas para seleccionar características basadas en un espacio latente multiómico. Como los repositorios de datos multiómicos sobre el cáncer siguen creciendo, la necesidad de estudiar y desarrollar enfoques multimodales será una prioridad en la genómica del cáncer.

Appendix B

Extended abstract in French

B.1 Introduction

Le rythme croissant de génération de données sur les profils tumoraux au cours de la dernière décennie a permis le développement d'algorithmes d'apprentissage statistique pour explorer et analyser le paysage des types et sous-types de tumeurs et la survie des patients d'un point de vue biomoléculaire. Les données tumorales sont généralement décrites par des caractéristiques trasncriptomiques et le niveau d'expression d'un gène-transcrit donné dans la cellule tumorale, ces caractéristiques peuvent donc être utilisées pour apprendre des règles statistiques qui améliorent la compréhension de l'état et du type d'une cellule cancéreuse.
Néanmoins, les données transcriptomiques des tumeurs sont très dimensionnelles et chaque tumeur peut être décrite par des milliers de caractéristiques génétiques, ce qui rend difficile l'apprentissage automatique et la compréhension des mécanismes biologiques sous-jacents. Cette thèse étudie comment réduire la dimensionnalité et gagner en interprétabilité sur les gènes qui codent le signal de la distribution des données en proposant des méthodes de réduction de dimension basées sur des pipelines de sélection et d'extraction de caractéristiques. Les méthodes proposées sont basées sur des modèles de variables latentes et des méthodes de noyau avec l'idée d'explorer la connexion entre les fonctions de similarité par paire d'échantillons de tumeurs et les espaces latents de faible dimension qui capturent la structure interne des données d'entraînement. Les méthodes proposées ont montré des améliorations dans les tâches de sélection de caractéristiques supervisées et non supervisées par rapport aux méthodes de référence pour classer et apprendre des sous-groupes de tumeurs respectivement. En outre, les méthodes proposées développées dans cette thèse ont montré leur adaptabilité pour traiter non seulement des données d'entrée mono-omiques mais aussi multi-omiques.

B.2 Chapitre 1

Le premier chapitre de cette thèse intitulé *Réduction de dimensionnalité sur les profils tumoraux biomédicaux : une approche d'apprentissage automatique* présente les défis et les

opportunités que la communauté de l'apprentissage automatique et de la reconnaissance des formes a pour l'exploration des maladies cancéreuses dans un contexte biomédical. Ces opportunités sont principalement propulsées par la croissance continue et la disponibilité de grandes données biomédicales et moléculaires provenant de profils tumoraux tels que l'ADN, l'ARN, les protéines et les métabolites [1]. De plus, d'un point de vue mathématique et statistique, les méthodes d'apprentissage automatique ont amélioré leurs capacités à traiter des données complexes et de grande dimension au cours des dernières décennies. Enfin, l'amélioration des processeurs de calcul permet la mise en œuvre efficace des méthodes d'apprentissage dans de grands ensembles de données [2].

B.2.1 Maladie cancéreuse

Le cancer est un groupe de maladies génétiques qui peuvent se développer dans n'importe quelle partie du corps humain. Il naît lorsque des cellules anciennes ou endommagées qui devraient mourir survivent et se développent de manière incontrôlée, altérant la fonction normale d'une cellule. Ces cellules commencent alors à se diviser sans s'arrêter pour former un amas de cellules appelé tumeur. La maladie cancéreuse survient à la suite des modifications et des mutations produites dans la séquence d'ADN du génome des cellules cancéreuses [4]. Le pire scénario est appelé *métastase* et il se produit lorsque la croissance des cellules tumorales d'un organe donné se poursuit, ce qui les force à se propager dans différents sites du corps, entraînant la défaillance de plusieurs organes et finalement la mort.
L'Organisation mondiale de la santé (OMS) estime à 9,6 millions le nombre de décès dus au cancer en 2018, soit un décès sur six, ce qui en fait la deuxième cause de décès dans le monde.

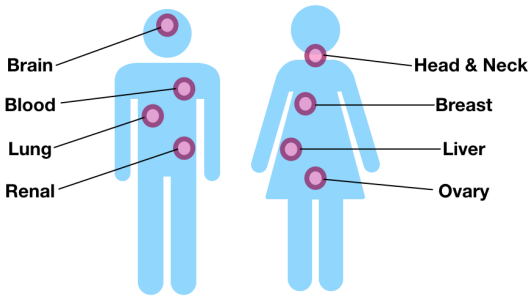

Figure B.1: Sites primaires des tumeurs

Bien que chaque tumeur puisse être caractérisée par le même ensemble de caractéristiques biologiques telles que l'ADN, l'ARN, les protéines ou les métabolites, le signal encodé dans ces derniers diffère considérablement en fonction du site primaire du corps où la tumeur est située, du stade de la tumeur et de son sous-type. Nous appelons ce signal codé "*signature génétique*" en référence à l'activité des différentes régions et parties du génome humain dans une cellule tumorale spécifique. Par conséquent, si d caractéristiques biologiques sont mesurées, elles peuvent être modélisées comme des variables aléatoires $x_1, x_2, ..., x_i, ..., x_d$

qui prennent des valeurs différentes dans différents échantillons de tumeurs, régies par une distribution de probabilité $p_1(x_1), p_2(x_2), ..., p_i(x_i), ..., p_d(x_d)$. Par exemple, une tumeur située dans le cerveau présentera une signature génétique différente de celle d'une tumeur située dans le foie. Même lorsque les tumeurs appartiennent au même type et au même site primaire, comme le rein, elles peuvent être classées en différents sous-types tels que le carcinome à cellules claires (KIRC) ou le carcinome papillaire (KIRP), chacun présentant une signature génétique différente associée à chaque sous-type. De plus, la signature génétique d'une tumeur de même type et sous-type peut différer en fonction du *pronostic* du patient, ceux qui ont un taux de survie faible ou élevé.
Il est également possible que le signal codé d'une signature génétique donnée soit situé dans une fraction seulement du génome humain. Cette situation révèle la nécessité de rechercher la région ou la zone du génome responsable du codage du signal afin de comprendre quels mécanismes biologiques sont impliqués.
Pour les raisons décrites ci-dessus, le traitement et l'interprétation des signatures génétiques des tumeurs est une tâche cruciale pour comprendre la complexité de la maladie cancéreuse et pour identifier avec précision une tumeur donnée par type, sous-type, stade ou pronostic. Il est important de noter que, contrairement à la prévention du cancer avant qu'il n'apparaisse en tant que maladie, cette thèse se concentre sur la caractérisation précise d'une tumeur une fois qu'elle existe déjà chez le patient. Le fait de pouvoir déterminer le type, le sous-type, le stade ou le pronostic associé d'une tumeur à partir d'une signature génétique permet de déterminer quel traitement doit être appliqué à chaque patient dans un contexte de médecine personnalisée. En outre, il est possible de découvrir des bio-marqueurs qui facilitent l'interprétation des données génomiques de la tumeur afin d'améliorer le processus de décision dans le diagnostic du cancer.

B.2.2 Données omiques sur les tumeurs

Chaque tumeur est caractérisée par différents groupes de caractéristiques moléculaires. Chacun de ces groupes de caractéristiques est composé du même type de macro-molécule et est appelé *omique*. Lorsqu'une tumeur est décrite simultanément par différentes omiques, le contexte est appelé *multi-omics*. [5]. Le contexte multi-omique décrit s'inscrit dans le concept de données multimodales, une situation couramment utilisée lorsque des caractéristiques sont obtenues à partir de sources et d'environnements différents, comme des images et le langage naturel par exemple [6]. Le contexte multi-omique du cancer implique des groupes de caractéristiques moléculaires telles que l'ADN, l'ARN, les protéines ou les métabolites, de sorte que chaque tumeur peut être décrite par de multiples omiques provenant de données biomoléculaires.
De plus, chaque tumeur a un phénotype, terme associé aux caractéristiques et aux traits observables. Le phénotype d'une tumeur est causé et déterminé à la fois par les caractéristiques omiques et par l'environnement. Étant donné que les maladies cancéreuses ont une forte composante génétique et multi-omique, cette thèse se concentre sur les caractéristiques omiques des tumeurs et sur la façon dont ces caractéristiques moléculaires peuvent déterminer un phénotype donné. Les phénotypes tumoraux peuvent être classés selon le site primaire où

la tumeur s'est développée en premier et selon le type histologique [8] qui est associé au type de tissu de la tumeur. En général, le site primaire définit le type de tumeur tandis que le tissu ou l'histologie définit le sous-type de tumeur. Le type et le sous-type de tumeur définissent le phénotype qui correspond à un diagnostic clinique posé par les médecins cliniciens. De plus, les données sur les tumeurs sont étiquetées par le phénotype mais aussi par des informations cliniques comme le stade de la tumeur, la survie du patient, son âge et son sexe.
Chaque omique contient des biomarqueurs potentiels qui pourraient aider à détecter un résultat tumoral comme le sous-type de la tumeur, le stade ou la survie du patient. Pour cette raison, la découverte de biomarqueurs est une tâche importante pour sélectionner les caractéristiques biomédicales importantes. En outre, la découverte de biomarqueurs permet aux médecins de se concentrer uniquement sur un ensemble réduit et restreint de caractéristiques au lieu de mesurer le génome complet d'un patient.

B.2.3 Données tumorales pour les tâches d'apprentissage automatique

Dans les sections précédentes, les données moléculaires omiques et les données cliniques ont été présentées et décrites. Dans cette section, l'objectif est d'expliquer comment les modèles d'apprentissage automatique peuvent être utilisés dans ce contexte pour estimer les résultats cliniques à partir des données omiques. Soit un espace d'entrée $\mathcal{X}$ un espace à d dimensions tel que $\mathcal{X} \subseteq \mathbb{R}^d$. Alors un ensemble S_u de n échantillons est défini dans $\mathcal{X}$ comme suit

$$S_u = \{x_1, x_2, ..., x_n\}$$

où chaque échantillon x_i est un vecteur de dimension d. Dans cette thèse, chaque vecteur échantillon est associé à un profil tumoral caractérisé par une donnée omique spécifique. Chaque ensemble de données omiques est composé de n profils tumoraux caractérisés par d caractéristiques génétiques. Si les informations cliniques sont disponibles, un vecteur d'étiquettes y peut être défini comme suit : $y \subseteq \mathbb{R}$ pour les variables cliniques continues comme les jours de survie, $y = \{-1, 1\}$ pour les étiquettes catégorielles binaires comme le sous-type de tumeur ou $y = \{1, 2, ..., T\}$ pour les étiquettes catégorielles multiples comme le site primaire ou le stade de la tumeur où T est le nombre d'étiquettes catégorielles cliniques. L'ensemble d'échantillons S_u peut alors être redéfini comme un ensemble S_s de paires entrée-sortie d'échantillons de tumeurs et d'étiquettes cliniques.

$$S_s = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$$

L'ensemble S_u correspond à un ensemble non supervisé d'échantillons puisqu'il n'y a pas d'étiquette clinique appariée aux échantillons alors que l'ensemble S_s est considéré comme un ensemble supervisé puisque chaque échantillon est apparié à une étiquette clinique. Dans cette thèse, je propose deux scénarios, celui où les étiquettes cliniques des tumeurs sont disponibles et celui où les étiquettes ne sont pas disponibles, appelés respectivement problèmes supervisés et non supervisés : [18]. La figure 2 montre un exemple d'un jeu de données sur les tumeurs

composé de mutations somatiques et d'expressions génétiques étiquetées avec des informations cliniques provenant de patients.

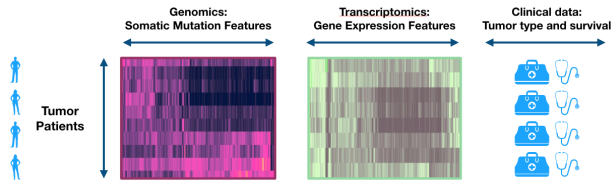

Figure B.2: Diagramme des données multi-omiques et cliniques

Dans le cas d'une approche supervisée, l'objectif est d'apprendre, à partir de l'ensemble S_s, une fonction $f(x)$ qui associe les échantillons de tumeurs x_i à une étiquette de tumeur y_i, telle que

$$f(x) \sim y \tag{B.1}$$

en minimisant une fonction de perte $L(y, f(x))$ qui mesure la divergence entre la sortie de $f(x)$ et les véritables étiquettes y. Lorsque les étiquettes sont continues $y \subseteq \mathbb{R}$, le problème est défini comme une régression. Lorsque les étiquettes sont catégoriques ou discrètes $y = \{-1, 1\}$, le problème est défini comme une classification.
Dans les cas de contextes non supervisés où les étiquettes de tumeurs y ne sont pas disponibles, le type de problèmes à résoudre est lié à l'apprentissage des similarités entre les échantillons de tumeurs de S_u et à la recherche de communautés ou de sous-groupes de tumeurs très similaires entre elles. Les groupes résultants sont appelés *clusters* et on s'attend à une corrélation entre ceux-ci et les étiquettes cliniques. La figure 3 montre un exemple de problèmes d'apprentissage supervisé et non supervisé.

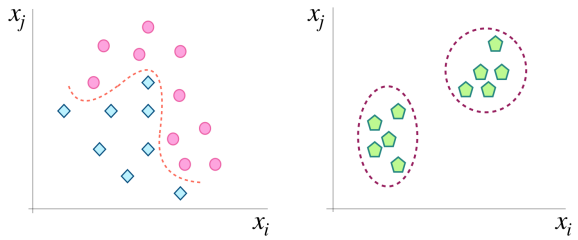

Figure B.3: Figure de gauche : un problème de classification supervisée. Figure de droite : un problème de regroupement non supervisé.

De plus, comme chaque profil de tumeur est caractérisé par plus de 20 000 gènes codant pour des protéines, l'espace dans lequel se trouvent les échantillons est très dimensionnel. Ainsi, une tâche importante d'apprentissage automatique à étudier dans cette thèse est la réduction de la dimensionnalité (*dimensionality reduction*). L'objectif de la réduction de

la dimensionnalité est de cartographier et de caractériser les données d'entrée dans une dimension inférieure. Pour un algorithme d'apprentissage automatique, les espaces de haute dimension et les échantillons de faible taille augmentent la complexité du problème, et l'apprentissage à partir de données dans ce contexte est moins efficace. Les ensembles de données omiques de tumeurs utilisés dans cette thèse ont un nombre de caractéristiques qui dépasse de loin le nombre d'échantillons. Cette situation est connue sous le nom de [**?**] et n'est pas souhaitable pour l'apprentissage statistique. De plus, il est supposé que l'espace d'entrée de haute dimension contient plusieurs caractéristiques bruyantes qui peuvent être éliminées. Ainsi, la dimensionnalité intrinsèque des données est supposée être inférieure à celle de l'entrée [20].
Pour cette raison, l'apprentissage de représentations de faible dimension des données d'entrée et la réduction de la dimensionnalité est une tâche importante étudiée dans cette thèse et il y a deux hypothèses importantes à considérer : premièrement, il n'est pas nécessaire d'apprendre à partir de l'ensemble complet des caractéristiques biologiques et un sous-ensemble réduit de gènes est juste nécessaire pour effectuer des tâches d'apprentissage automatique supervisé ou non supervisé. Deuxièmement, il existe des représentations latentes plus simples et de faible dimension des profils tumoraux qui aident à comprendre la distribution et la structure sous-jacentes des données.

B.2.4 Challenges to be addressed in this thesis

Les données omiques et cliniques des tumeurs présentées invitent à étudier et à développer des modèles d'apprentissage automatique pour évaluer différents problèmes. Les informations cliniques telles que le sous-type de tumeur, le stade de la tumeur ou les jours de survie peuvent être estimées à partir des données omiques dans une approche supervisée. En outre, des approches non supervisées sont également nécessaires car les étiquettes cliniques ne sont pas toujours disponibles. Il est donc possible de découvrir des groupes de tumeurs qui présentent une grande similarité entre les profils de tumeurs.
Que ce soit pour les approches supervisées ou non supervisées, les données omiques utilisées sont de haute dimension et le nombre d'échantillons de tumeurs est faible. Ce scénario de haute dimension rend difficile l'exécution de tâches supervisées ou non supervisées. Il existe des caractéristiques génétiques plus informatives que d'autres dans les ensembles de données sur les mutations somatiques et l'expression génique. C'est pourquoi il est nécessaire de développer des méthodes robustes de réduction de la dimensionnalité pour extraire des informations utiles des données omiques. La réduction de la dimensionnalité peut être effectuée en sélectionnant un sous-ensemble de gènes p où $p << d$ ou en apprenant une nouvelle représentation de faible dimensionnalité en combinant les caractéristiques d'entrée et en filtrant le bruit d'entrée.
Cette thèse n'utilisera les données d'expression génique que comme un seul scénario omique tandis que le dernier chapitre de la thèse étudie la fusion de données multi-omiques qui et comment combiner plusieurs omiques tels que la mutation somatique et l'expression génique.

B.3 Méthodes à noyaux et sélection supervisée de caractéristiques

Le paysage des profils tumoraux *omique* est composé et catégorisé par un large éventail de classes cliniques différentes telles que les types de tumeurs, les sous-types et la survie des patients, ce qui fait de la classification des profils tumoraux une tâche clé pour l'apprentissage automatique dans l'informatique du cancer. La classification des tumeurs implique de trouver une fonction de décision mathématique utilisée comme frontière de décision entre les classes de tumeurs dans un espace de haute dimension $\mathcal{X}$ défini par les caractéristiques d'expression génique d'entrée qui sont considérées comme des variables aléatoires. L'idée centrale de la classification est d'apprendre la fonction de décision de manière à attribuer des étiquettes cliniques précises à des profils tumoraux dont le contexte clinique est inconnu. Le problème de la classification des profils tumoraux à l'aide de données d'expression génétique est de haute dimension puisqu'il est caractérisé par des dizaines de milliers de caractéristiques d'expression génétique, comme nous l'avons présenté dans le chapitre 01. De plus, les ensembles de données sur le cancer ont généralement une taille d'échantillon faible. Comme mentionné dans le chapitre précédent, le rapport entre la taille de l'échantillon n et la dimension de l'espace des caractéristiques d de ces jeux de données est $\frac{n}{d}$. $<< 1$ ce qui caractérise les situations où la malédiction de la dimensionnalité est un problème de première importance. Pour s'attaquer à un problème de classification de tumeurs dans un contexte de haute dimensionnalité, réduire la dimensionnalité des données est une tâche importante à réaliser. La sélection des caractéristiques est une famille de méthodes de réduction de la dimensionnalité conçues pour écarter les caractéristiques bruyantes et inutiles tout en ne conservant qu'un sous-ensemble des caractéristiques d'entrée utiles pour effectuer une tâche d'apprentissage automatique supervisée ou non supervisée. On suppose que les caractéristiques sélectionnées contiennent un signal pertinent pour une tâche d'apprentissage supervisé telle que la classification, tandis que les caractéristiques écartées contiennent du bruit ou un signal redondant. Un autre objectif important des méthodes de sélection de caractéristiques est de guider le processus de découverte de biomarqueurs [23]. [24]. Les tumeurs étant caractérisées par des dizaines de milliers de caractéristiques, il est coûteux et inefficace de mesurer toutes les caractéristiques biologiques. De plus, la sélection d'un sous-ensemble de caractéristiques génétiques peut améliorer une tâche d'apprentissage supervisé, comme la classification des tumeurs ou la régression de la survie, et améliore également l'interprétation biomédicale puisque les caractéristiques sélectionnées permettent de comprendre le système biologique sous-jacent. La sélection de caractéristiques a été largement appliquée en bioinformatique et en génomique du cancer à partir de données d'expression génique sur des problèmes d'apprentissage supervisé [25]. [26] [27]. L'hypothèse soutenue dans ce travail stipule qu'un sous-ensemble réduit de gènes est nécessaire pour effectuer et améliorer significativement la tâche d'apprentissage supervisé, en particulier dans ce travail, la tâche supervisée est la classification des tumeurs. La tâche de classification en génomique du cancer prend en entrée un profil tumoral d'expression génique donné et lui attribue une étiquette phénotypique en mettant en œuvre des algorithmes qui prennent des décisions en fonction de l'optimisation d'une fonction de

perte. Ces algorithmes apprennent des règles statistiques à partir de données étiquetées afin de construire une fonction de décision.

Les méthodes d'apprentissage supervisé les plus répandues, telles que les machines à vecteurs de support et les perceptrons, ou les méthodes non supervisées, telles que l'analyse en composantes principales, utilisent un produit interne (produit interne) entre les échantillons de données pendant le processus d'apprentissage, qui peut être interprété comme une mesure de similarité entre l'échantillon x_i et l'échantillon x_j (produit interne). Bien que cette mesure permette l'utilisation d'une famille de fonctions linéaires, elle peut être restreinte et limitée dans les applications du monde réel comme les données de tumeurs omiques de haute dimension. Afin d'étendre les capacités de ces méthodes, l'introduction de méthodes à noyau permet d'utiliser ce cadre linéaire pour traiter des fonctions non linéaires plus polyvalentes. Leur étude est au cœur de cette thèse [18]. Les noyaux ont été largement utilisés en biologie informatique pour des tâches supervisées comme la classification ou la régression et pour des tâches non supervisées comme le clustering ainsi que pour résoudre des problèmes de réduction de dimensionnalité [29]. [30]. On peut donc s'attendre à ce que les méthodes à noyaux puissent également améliorer la tâche d'apprentissage sur les données omiques.
Dans ce contexte, une série de questions se pose : étant donné un profil d'expression génétique, comment pouvons-nous classer des échantillons de tumeurs sur la base d'attributs cliniques ? Si les données sont de haute dimension, comment peuvent-elles être réduites à un espace de basse dimension facile à traiter ? Si les méthodes linéaires peuvent être limitées dans les données tumorales de haute dimension, comment les méthodes non linéaires peuvent-elles être utilisées pour améliorer la tâche d'apprentissage ? Dans ce chapitre, l'apprentissage supervisé est utilisé pour la classification des tumeurs à partir d'un profil d'expression génétique. La sélection des caractéristiques est utilisée pour sélectionner un sous-ensemble de gènes qui améliore la performance de la classification et qui aide à interpréter la signature biologique impliquée dans la tâche de classification. Les méthodes à noyaux sont utilisées pour effectuer et améliorer à la fois la classification des tumeurs et les méthodes de sélection des caractéristiques. Elles sont à la base d'une nouvelle approche proposée pour les tâches de sélection des caractéristiques sur les données d'expression génique.
Ce chapitre présente d'abord dans l'ordre suivant les algorithmes de classification, les méthodes à noyau et les modèles de sélection de caractéristiques. La section sur la classification définit les classificateurs linéaires, la méthode de classification par machine à vecteurs de support et les mesures d'évaluation de la classification. La section sur les méthodes à noyaux présente la métrique d'alignement des noyaux, l'apprentissage à noyaux multiples, l'astuce des noyaux et la façon dont les noyaux peuvent être utilisés pour traiter les problèmes de classification où les classes ne sont pas linéairement séparables. La section suivante porte sur l'approche de la sélection des caractéristiques, les méthodes de référence sont présentées et les mesures de performance sont définies. Enfin, une nouvelle approche appelée *Kernel Latent Regularization Feature Selection*(KLR-FS) [31] est proposée et les résultats expérimentaux sont détaillés.

B.3.1 Méthode proposée : Sélection des caractéristiques par régularisation latente à noyau

La plupart du temps, les caractéristiques importantes sont sélectionnées à l'aide d'une fonction objective supervisée [57]. Dans certains cas, comme le manque de données, l'objectif supervisé peut être trop strict et difficile à atteindre afin d'obtenir un modèle qui pourrait généraliser sur de nouvelles données non vues [58]. Dans ce scénario limité, l'idée proposée consiste à considérer la structure des données comme une autre source d'apprentissage en plus des étiquettes de l'échantillon y et à l'utiliser comme stratégie d'amélioration pour les tâches de sélection de caractéristiques et de classification. Une question majeure se pose alors : est-il possible d'améliorer le processus de sélection de caractéristiques en contraignant le sous-espace de caractéristiques sélectionné non seulement pour déduire les étiquettes ciblées mais aussi pour conserver plus d'informations sur la structure des données dans l'espace de caractéristiques initial ?

Ce travail propose une méthode de sélection de caractéristiques basée sur l'apprentissage à noyaux multiples (MKL). De plus, la méthode proposée combine MKL et un modèle d'extraction de caractéristiques latentes non linéaires pour améliorer le processus de sélection de caractéristiques par une combinaison d'approches supervisées et non supervisées respectivement. Cette combinaison d'approches vise à améliorer la capacité de généralisation de la classification des caractéristiques sélectionnées. Cette stratégie vise à maximiser la séparabilité entre les classes de tumeurs tout en considérant simultanément la structure latente des données d'entraînement. La méthode de sélection proposée effectue ce que nous appelons une *régularisation latente* en utilisant simultanément les étiquettes des données et les variables latentes non supervisées. Pour extraire les variables latentes des données d'apprentissage, un modèle de réduction de la dimensionnalité non supervisé est utilisé, plus précisément l'ACP à noyau (ACPK) [59]. L'idée principale est de rechercher des caractéristiques qui tiennent compte des étiquettes des tumeurs et préservent simultanément la structure des données. L'espace obtenu devrait être plus robuste et conduire à une meilleure généralisation lors des tâches de classification.

La méthode proposée est conçue pour traiter les problèmes de classification des tumeurs où la dimensionnalité $d > 18.000$ et la taille de l'échantillon est inférieure à $n < 200$ profils tumoraux. Les profils tumoraux sont classés par stade ou par pronostic. Dans ce scénario, la plupart des algorithmes de sélection de caractéristiques peuvent être sous-performants en raison du manque d'échantillons de tumeurs et de l'ensemble de caractéristiques de haute dimension. La régularisation latente proposée fonctionne comme un processus de relaxation des étiquettes qui améliore les performances de la classification des tumeurs sur de nouveaux échantillons de test non vus. La méthode proposée est appelée *Kernel Latent Regularization Feature Selection* (KLR-FS).

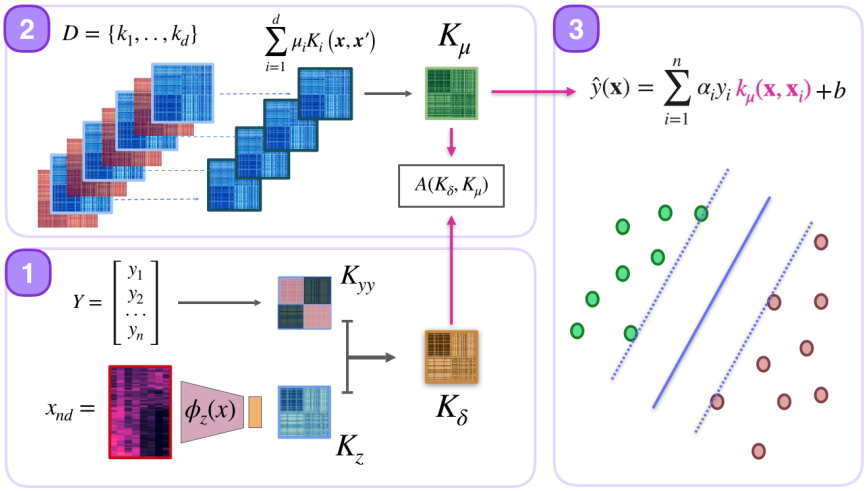

Figure B.4: Le pipeline KLR-FS. 1) Une matrice K_{yy} est construite à l'aide des étiquettes de tumeurs. Un modèle ACP à noyau est formé à l'aide des données de formation et un noyau K_z est construit sur l'espace latent Z appris à partir de l'ACP à noyau. La matrice du noyau K_δ est obtenue à partir du mélange entre les matrices K_{yy} et K_z. 2) À partir des données d'apprentissage, un ensemble D de noyaux à caractéristiques d est construit. Par MKL, un sous-ensemble de noyaux de caractéristiques p est sélectionné et une matrice de noyau K_μ est obtenue en améliorant l'alignement avec le noyau K_δ. Le vecteur μ indique les caractéristiques sélectionnées. 3) La fonction noyau k_μ est utilisée dans la classification par vecteurs de support..

B.3.2 Résultats et discussion

Ce travail propose un modèle de sélection de caractéristiques basé sur l'apprentissage à noyaux multiples couplé à l'ACP à noyaux avec une application à la classification des tumeurs en utilisant des données d'expression génique de haute dimension provenant de profils tumoraux. La méthode proposée KLR-FS vise à sélectionner les caractéristiques en tenant compte non seulement des étiquettes d'échantillon y mais aussi des variables latentes z des données d'entraînement présentées dans ce travail comme une régularisation *latent* puisqu'elle est utilisée pour améliorer la capacité de généralisation dans la sélection des caractéristiques pour la classification. L'idée est basée sur la considération non seulement d'un noyau cible supervisé pur k_{yy} mais aussi d'un noyau k_z construit à partir des variables latentes z obtenues à partir d'une réduction de dimensionnalité non linéaire ϕ_z. La régularisation latente provient d'un noyau construit sur l'espace latent $\mathcal{Z}$ généré par l'ACP à noyau. Cela permet d'explorer la relaxation des étiquettes de l'échantillon par une combinaison linéaire d'un noyau supervisé et d'un noyau non supervisé. Les expériences montrent que la meilleure performance de classification et le plus faible taux de redondance dans la sélection des caractéristiques pour le KLR-FS sont obtenus lorsque l'information latente est mélangée à l'information ciblée sur les étiquettes. Ces résultats montrent que la sélection de variables tenant compte à la fois de la structure de l'espace latent $\mathcal{Z}$ des données et des étiquettes du problème super-

visé peut améliorer la tâche de classification. De plus, l'apprentissage partiel d'un espace latent fonctionne comme un terme de régularisation puisqu'il limite l'espace de solution en introduisant la nécessité de capturer également la structure générale des données. Ce faisant, il peut améliorer la capacité de généralisation sur de nouveaux échantillons de test non vus. En outre, il existe une relation entre la structure latente des données et les étiquettes d'échantillons y et le fait d'apprendre uniquement à partir des étiquettes avec un critère supervisé classique peut entraîner un surajustement lors de l'apprentissage dans un contexte où le nombre d'échantillons est très limité. Par conséquent, la sélection de caractéristiques avec un mélange de critères latents supervisés et non supervisés peut diminuer le biais d'un classificateur en aval.
KLR-FS produit un noyau personnalisé k_μ qui est utilisé dans la classification par vecteur de support et contribue à améliorer les performances de classification. Le noyau k_μ résultant peut être utilisé non seulement pour les tâches de classification, mais aussi pour la représentation des profils tumoraux ou la visualisation par ACP à noyau, car il est construit à partir de sources latentes supervisées et non supervisées.
La méthode proposée est pertinente dans les problèmes où la structure des données d'apprentissage est partiellement corrélée aux étiquettes des tumeurs. Ainsi, l'apprentissage à partir de la structure latente des données améliore la tâche d'apprentissage supervisé.

B.4 Régularisation latente à noyau sur les méthodes de réduction suffisante pour l'analyse de survie

La réduction de dimension est un domaine important de problèmes dans la statistique contemporaine et l'apprentissage automatique. Elle vise à éviter la malédiction de la dimensionnalité dans les tâches d'apprentissage en aval. Comme expliqué dans les chapitres précédents, la réduction de la dimensionnalité joue un rôle clé dans l'analyse du profil des tumeurs. Le chapitre 01 a détaillé comment une seule tumeur peut être caractérisée par des dizaines de milliers de caractéristiques d'expression génique et l'importance de la réduction de la dimensionnalité. Le chapitre 02 étudie la sélection supervisée de caractéristiques pour réduire la dimensionnalité en conservant l'interprétabilité de la réduction et pour effectuer une tâche d'apprentissage en aval comme la classification.
En particulier, le chapitre 02 propose la méthode de sélection de caractéristiques par régularisation latente du noyau (KLR-FS). Le terme de régularisation latente dans KLR-FS vise à mélanger les matrices de noyau supervisées et non supervisées pour construire un noyau cible hybride qui guide le processus de sélection des caractéristiques. Par conséquent, les caractéristiques sélectionnées sont celles qui capturent des informations sur les étiquettes mais aussi la structure latente $\mathcal{Z}$ des données d'apprentissage. Cette régularisation latente montre comment la tâche d'apprentissage en aval, comme la classification, est améliorée lorsque le processus d'apprentissage prend en compte simultanément les étiquettes de la tumeur et la structure latente non supervisée. Dans ce chapitre, je considère intéressant d'explorer la généralisation du concept de régularisation latente au-delà de la méthode KLR-FS. L'une

des caractéristiques de la régularisation latente à noyau réside dans le fait qu'elle peut être utilisée pour toute tâche d'apprentissage supervisé ayant une fonction noyau cible pendant le processus d'apprentissage. Pour cette raison, ce chapitre vise à étendre le concept de régularisation latente à d'autres méthodes d'apprentissage qui utilisent des noyaux comme fonctions cibles.
La réduction de la dimensionnalité est couramment effectuée par des approches d'apprentissage non supervisé et est une tâche qui peut être effectuée par l'ACP [73], l'ACP à noyau [74], les autoencodeurs [75], t-SNE [76] entre autres. Ces méthodes apprennent un espace latent de faible dimension $\mathcal{Z}$ et la fonction objectif est conçue pour optimiser la préservation de la structure d'une distribution d'entrée X d'intérêt. D'autre part, les approches d'apprentissage supervisé pour la réduction de la dimensionnalité sont axées sur la préservation de l'information sur une réponse cible y [77]. Dans ce chapitre, l'approche de réduction suffisante (DRS) est particulièrement étudiée [78]. La RDS concerne la famille des méthodes supervisées de réduction de la dimensionnalité utilisées pour la classification et la régression. En tant qu'approche supervisée, la DTS considère les étiquettes d'échantillon y pour apprendre une projection $\rho(x)$ des données d'entrée $X \in \mathbb{R}^d$ dans un espace de faible dimension $\mathcal{S} \in \mathbb{R}^p$ où $p < d$ tel que [79]

$$\rho(\mathcal{X}) = \mathcal{S} \tag{B.2}$$

Formellement, le sous-espace $\mathcal{S}$ est un *sous-espace de réduction de dimension* en termes d'indépendance conditionnelle suivante

$$Y \perp X \mid \beta X \tag{B.3}$$

où $\perp$ indique l'indépendance et βX représente la projection orthogonale de X dans $\mathcal{S}$. La famille des méthodes SDR peut être utile pour prédire les étiquettes de tumeurs comme le pronostic [80]. Les méthodes SDR peuvent être utilisées pour apprendre un sous-espace $\mathcal{S}$ qui capture suffisamment d'informations sur la relation entre les caractéristiques d'expression des gènes et le résultat du phénotype tel que la survie du patient [81].
Une caractéristique importante des méthodes SDR est qu'elles ne considèrent que Y pour estimer le sous-espace $\mathcal{S}$ et que la structure non supervisée des données n'est pas utilisée. Ce scénario peut entraîner un surajustement dans un contexte où le nombre de caractéristiques d'expression génique d est supérieur au nombre d'échantillons de tumeurs n car $n < d$. L'hypothèse étudiée repose sur l'idée que les tâches d'apprentissage supervisé comme la réduction de la dimensionnalité peuvent être améliorées en apprenant simultanément à partir des étiquettes des échantillons et de la structure latente des données d'entrée. Ce chapitre vise à répondre à la question suivante : est-il possible d'améliorer une tâche d'apprentissage par régression couplée à des réductions de dimensionnalité suffisantes en introduisant la régularisation latente ? Comme les approches de la RDS sont conçues pour la classification et la régression, ce chapitre étudie comment la RDS peut être utilisée pour l'analyse de la survie des patients en apprenant un espace supervisé de faible dimension. La proposition de ce chapitre est une méthode qui améliore la régression de survie effectuée sur l'espace projeté de

la DTS $\mathcal{S}$ en utilisant une méthode de régularisation latente à noyau qui apprend à partir des étiquettes de tumeurs et de la structure latente non supervisée des données d'apprentissage [31].

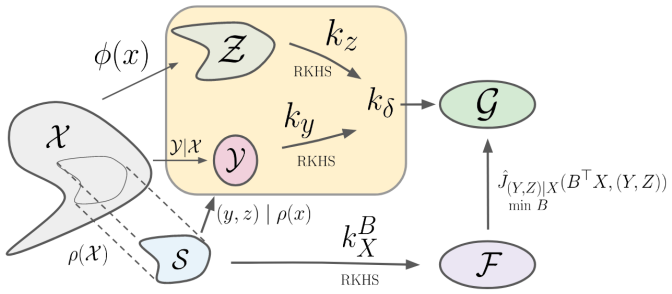

Figure B.5: Méthode KLR-SDR proposée. Le noyau hybride k_δ mélange les informations supervisées et non supervisées provenant respectivement de k_y et k_z. Le sous-espace de réduction obtenu, $\mathcal{S}$, explique à la fois les étiquettes et la structure latente des données de transfert.

La figure 5 montre un diagramme de la méthode proposée, appelée *Régularisation latente par noyau - Réduction dimensionnelle suffisante* (KLR-SDR). La principale contribution de ce chapitre est basée sur l'extension du concept de régularisation latente de noyau proposé dans le chapitre 2 intitulé *Régularisation latente de noyau pour la sélection de caractéristiques dans la classification des tumeurs.* [31] à la méthode Kernel Sufficient Dimension Reduction (KSDR) [78]. Le chapitre de la thèse propose des expériences sur des données jouets synthétiques pour la classification et la régression de survie pour des données réelles d'expression génique de tumeurs. Dans les deux cas, un espace latent supervisé de faible dimension est appris, puis les tâches d'apprentissage supervisé telles que la classification et la régression sont mises en œuvre.

B.4.1 Résultats et discussion

Les méthodes SDR sont des méthodes de réduction de dimension adaptées aux tâches d'apprentissage supervisé. En particulier, l'analyse de régression implique un type de modèles capables de traiter les problèmes d'analyse de survie. C'est pourquoi les méthodes de DTS sont adaptées au traitement des données de survie et d'expression génétique des patients atteints de cancer du Consortium international sur le génome du cancer (International Cancer Genome Consortium [90]).
La motivation pour utiliser la régularisation latente à noyau est basée sur l'idée de traiter une quantité limitée de données (petit n), comme un moyen de préserver la structure de la classe interne et de tenir compte de l'ignorance du contenu informationnel du jeu de données initial. Les méthodes SDR standard ne parviennent pas à conserver la distribution structurelle des données après réduction de la dimension et ne capturent que l'information

discriminante de la distribution conditionnelle $X \mid Y$. Cette situation entraîne, d'une part, l'incapacité d'explorer les sous-groupes au sein d'une classe de tumeurs après réduction de la DTS et, d'autre part, un surajustement puisque l'apprentissage d'une représentation de faible dimension uniquement à partir des étiquettes dans un contexte de faible taille d'échantillon peut limiter la représentation obtenue à de nouveaux échantillons de test non vus, manquant ainsi la capacité de généralisation.
La méthode SDR à noyau standard utilise un noyau cible K_y pour guider la réduction. Dans ce travail, il est proposé de le remplacer par le noyau K_δ comme cible afin d'effectuer la réduction de dimensionnalité d'une approche supervisée à une approche non supervisée dans le même cadre. En outre, les expériences sur des données synthétiques pour la classification binaire révèlent comment la performance d'un classificateur à vecteur de support est améliorée lorsque la régularisation latente à noyau est utilisée dans la tâche de réduction de dimension. Des expériences utilisant des données réelles de profils d'expression génique de patients cancéreux étiquetés avec des informations de survie montrent comment apprendre une représentation de faible dimension $\mathcal{S}$ régularisée par une régularisation latente à noyau qui est finalement utilisée pour entraîner un modèle de régression de Cox pour la prédiction de survie. La régularisation latente à noyau améliore la prédiction de la survie dans tous les ensembles de données par rapport au cas entièrement supervisé. La méthode proposée de réduction de dimension suffisante par régularisation latente à noyau (KLR-SDR) avec le paramètre de mélange δ qui maximise la performance de la régression est comparée aux méthodes kSDR, PCA et kPCA. La méthode KLR-SDR est plus performante que toutes les méthodes de référence, ce qui montre le potentiel du module de régularisation latente. Dans ce travail, la méthode kernel-PCA est utilisée pour extraire un espace latent $\mathcal{Z}$ pour le terme de régularisation latent, néanmoins toute autre méthode non supervisée peut être utilisée telle que t-SNE, Autoencodeurs ou PCA.
Les résultats obtenus dans des données synthétiques et réelles montrent que l'apprentissage non seulement à partir d'étiquettes supervisées mais aussi à partir d'une structure latente non supervisée améliore la tâche d'apprentissage supervisé en aval. Les résultats suggèrent que la régularisation latente améliore à la fois les tâches de classification et de régression ; pour cette raison, l'approche de régularisation latente à noyau proposée pourrait être généralisée à toute méthode d'apprentissage supervisé qui utilise des noyaux pour guider le processus d'apprentissage.
La régularisation Kernel Latent améliore une méthode de réduction de dimension Kernel Sufficient pour les problèmes de régression de survie des patients atteints de cancer. Les résultats suggèrent que l'apprentissage à partir de la structure latente des données de formation améliore les tâches d'apprentissage supervisé. La motivation de l'utilisation de la structure interne des données pendant l'apprentissage est basée sur la prise en compte de l'ignorance du contenu informationnel de l'ensemble des données initiales puisque la faible densité des données peut ne pas permettre la découverte de sous-clusters de classes internes. De plus, l'utilisation d'une structure de données mélangée à des étiquettes améliore la généralisation d'une tâche supervisée sur de nouvelles données de test non vues.
Les méthodes à noyaux ont montré leur potentiel d'amélioration même dans des scénarios

complexes où le nombre de caractéristiques dépasse de loin le nombre d'échantillons disponibles $d > n$ comme les données génomiques du cancer. Les travaux futurs porteront sur les données multi-omiques des patients.

B.5 Sélection non supervisée de caractéristiques avec des autoencodeurs et des méthodes à noyau

Ce chapitre est consacré à l'exploration et à la compréhension des méthodes de sélection de caractéristiques et de réduction de la dimensionnalité à partir de données d'expression génique de profils tumoraux dans le but de découvrir des biomarqueurs et des sous-types de tumeurs via des méthodes d'apprentissage automatique non supervisées. Le chapitre 02 définit les problèmes d'apprentissage supervisé et les méthodes de sélection des caractéristiques pour améliorer une tâche de classification. Un aspect principal du chapitre 02 est la disponibilité des étiquettes de tumeur y comme le sous-type ou la survie pendant l'apprentissage du modèle qui permet l'application de classifieurs comme les machines à vecteurs de support. De plus, dans le chapitre 02, les étiquettes de tumeurs sont utilisées pendant le processus de sélection des caractéristiques, ainsi non seulement la classification mais aussi la sélection sont supervisées. Néanmoins, les étiquettes de tumeurs ne sont pas toujours disponibles au cours du processus d'apprentissage et des méthodes non supervisées sont nécessaires dans ce cas [26]. L'apprentissage non supervisé est une vaste branche des méthodes d'apprentissage automatique qui va de la réduction de la dimensionnalité au regroupement. De plus, le contexte hautement dimensionnel des données génomiques des profils tumoraux nécessite de réduire la dimensionnalité pour améliorer la performance d'une tâche d'apprentissage non supervisé en aval. L'une des questions auxquelles ce chapitre tente de répondre est la suivante : est-il possible de sélectionner des gènes sans étiquette clinique de manière à faciliter la découverte de sous-types de tumeurs ? Ce chapitre formalise et explique comment sélectionner un sous-ensemble réduit de gènes qui améliore le regroupement des tumeurs afin de guider la découverte de nouveaux sous-types de tumeurs [96]. L'interprétation du système biologique se fait par la découverte de biomarqueurs avec des méthodes de sélection de caractéristiques et la découverte de sous-types par une tâche de regroupement.
Chaque profil de tumeur est décrit par plus de $d = 10.000$ caractéristiques d'expression génique par rapport à une taille d'échantillon relativement faible n, ce qui définit un espace d'entrée de grande dimension et une complexité supplémentaire dans l'analyse des données. En outre, les types de tumeurs présentent une hétérogénéité interne qui peut être subdivisée en sous-groupes présentant des caractéristiques cliniques communes et appelés sous-types de tumeurs. La découverte de sous-types de tumeurs est donc une tâche importante en génomique du cancer, car elle permet aux médecins cliniciens d'attribuer un diagnostic et des traitements différents, même au sein d'un même type de tumeur [97] [98]. [99]. Étant donné l'espace de haute dimension des données d'entrée, il est nécessaire de réduire la dimensionnalité tout en préservant l'interprétation biologique du système comme une signature génique réduite. Ces raisons motivent ce chapitre à proposer deux nouvelles méthodes non supervisées de sélection

de caractéristiques pour le regroupement de tumeurs.
Comme nous l'avons expliqué au chapitre 02, les méthodes de sélection de caractéristiques sont nécessaires en génomique du cancer, car elles fournissent une représentation de faible dimension des données d'entrée, caractérisée par un sous-ensemble sélectionné de gènes d'entrée permettant d'interpréter les résultats et d'éliminer le reste en suivant une fonction objective liée à l'amélioration d'une tâche d'apprentissage [100]. En outre, les gènes sélectionnés peuvent être utilisés pour guider les stratégies de découverte de biomarqueurs [24]. Le sous-espace résultant obtenu par une méthode de sélection de caractéristiques est décrit explicitement par un sous-ensemble de caractéristiques biologiques en supposant que l'ensemble de caractéristiques initial contient des caractéristiques bruitées qui peuvent être écartées. Un sous-ensemble de caractéristiques réduit a pour avantage de réduire la complexité du modèle et de ne mesurer qu'un ensemble réduit de biomarqueurs [26].
Dans un problème non supervisé, puisque les étiquettes y ne sont pas disponibles pendant la formation, on s'attend à ce que les gènes sélectionnés révèlent la structure des données, ce qui est évalué par une amélioration du regroupement des profils tumoraux. Pour guider le processus de sélection des caractéristiques, ce travail propose une méthode qui apprend d'abord une représentation de faible dimension et débruitée des données d'entrée, connue sous le nom d'espace latent $\mathcal{Z}$, qui est apprise par un réseau neuronal non supervisé connu sous le nom d'auto-encodeur. Deux stratégies sont ensuite proposées dans ce chapitre. La première consiste à utiliser un modèle d'apprentissage à noyaux multiples pour sélectionner un sous-ensemble de caractéristiques génétiques dans le but d'aligner autant que possible la distribution résultante des caractéristiques sélectionnées sur une distribution cible des échantillons d'apprentissage dans la représentation apprise $\mathcal{Z}$. La deuxième stratégie, au lieu d'utiliser une méthode d'apprentissage à noyaux multiples pour la sélection des caractéristiques, utilise un autoencodeur vanille comme réseau d'apprentissage pour apprendre un deuxième espace latent de faible dimension $\mathcal{Z}_v$ de telle sorte que la divergence de la distribution des données dans $\mathcal{Z}_v$ par rapport à celle de l'espace latent $\mathcal{Z}$ soit minimisée en utilisant la distance de divergence moyenne maximale (MMD) [101]. Enfin, dans les deux méthodes proposées, le regroupement sur le sous-espace obtenu à partir des caractéristiques sélectionnées devrait permettre d'observer les attributs cliniques significatifs associés à chaque groupe et de valider ainsi la qualité des caractéristiques sélectionnées.
La principale contribution de ce chapitre concerne deux méthodes non supervisées capables de sélectionner des gènes ayant une pertinence clinique à partir de données d'expression génique de haute dimension, sans avoir besoin d'étiquettes de tumeurs.

B.5.1 Apprentissage non supervisé

L'objectif de l'apprentissage non supervisé est de capturer la structure de la distribution des données en se basant sur la similarité entre les échantillons de données. La principale différence entre l'apprentissage supervisé et l'apprentissage non supervisé est que ce dernier n'utilise aucune étiquette de réponse y pendant l'apprentissage d'un modèle. Dans cette ligne, la sélection de caractéristiques non supervisée repose sur la recherche d'un sous-ensemble

de caractéristiques d'entrée qui capture aussi bien la structure des données sans utiliser d'étiquette de réponse.
Dans le contexte de la génomique du cancer, les données d'expression génétique sont supposées être une variable aléatoire indépendante représentée par une matrice d'entrée $\mathbf{X}$. La matrice d'entrée $\mathbf{X}$ représente un ensemble de n vecteurs $d-$ d'entrée x_i, $i = 1...N$ et x_i est le $i$ème échantillon de l'ensemble de données. Chaque vecteur d'entrée x_i est associé à un échantillon de tumeur et chaque dimension j du vecteur d'entrée x_{ij} est un gène et représente le niveau d'expression correspondant. Les vecteurs d'entrée sont des échantillons aléatoires à valeur réelle obtenus à partir d'un espace $d-$dimensionnel $\mathcal{X} \in \mathbb{R}^d$. Alors, étant donné un ensemble S de n échantillons

$$S = \{x_1, x_2, ..., x_n\} \tag{B.4}$$

l'objectif de l'apprentissage non supervisé est de décrire la structure, la distribution des données et la similarité entre les échantillons de l'ensemble S.
La motivation de l'utilisation de l'apprentissage non supervisé pour la sélection des caractéristiques des profils tumoraux repose sur deux hypothèses. La première hypothèse stipule que dans une classe de type de tumeur connue, il peut exister des sous-groupes internes et une hétérogénéité interne d'importance clinique. Il est donc supposé qu'un ensemble de données sur les tumeurs peut présenter des groupes ou des grappes d'échantillons de tumeurs avec une intra similarité élevée qui ne sont pas décrits par une étiquette clinique, de sorte que la recherche de ces sous-groupes est souhaitable. La deuxième hypothèse repose sur l'idée que la distribution et la structure des données d'entrée ont une *dimensionnalité intrinsèque* p inférieure à la dimensionnalité initiale d'entrée d. Pour cette raison, on suppose qu'il est possible de décrire la variabilité et la structure des données d'apprentissage dans une dimension inférieure.
Dans ce travail, les approches de *clustering* [102], *modèles à variables latentes* [60] et *méthodes à noyau*. [18] sont utilisées pour résoudre les problèmes de recherche de sous-groupes et réduire la dimensionnalité des données d'entrée. Ces méthodes sont combinées pour effectuer une sélection non supervisée de caractéristiques sur des données d'expression génique provenant de profils tumoraux.

B.5.2 Méthodes proposées

Pour effectuer une sélection non supervisée des caractéristiques, deux méthodes sont présentées dans ce chapitre. La première est la sélection de caractéristiques par noyau latent. Elle est composée de deux modèles : un autoencodeur et une méthode d'apprentissage à noyaux multiples. Le premier, l'auto-codeur, est construit pour apprendre un espace latent de faible dimension. Ensuite, la tâche d'apprentissage à noyaux multiples est conçue pour sélectionner les caractéristiques qui suivent la distribution des données de l'espace latent, comme le montre la figure B6.

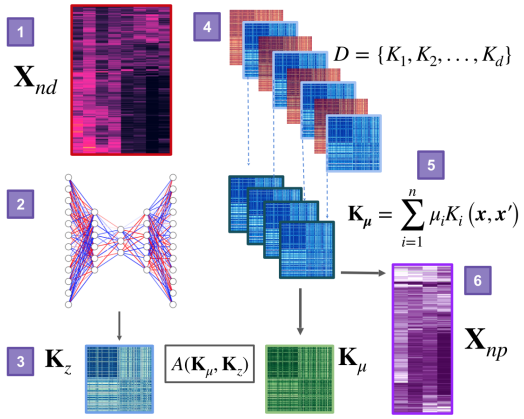

Figure B.6: Taille de la matrice de données $\mathbf{X}$ avant (gauche) et après (droite) la sélection des caractéristiques. d est le nombre de caractéristiques d'entrée et p est le nombre de caractéristiques sélectionnées..

La deuxième méthode proposée dans ce chapitre est basée sur deux autoencodeurs et est appelée Latent Maximum Mean Discrepancy Feature Selection (LMMD-FS). Le premier est chargé d'apprendre un espace latent de faible dimension des données d'entrée comme le fait la méthode LKFS. Le deuxième auto-codeur est un auto-codeur vanille à une seule couche chargé de sélectionner les caractéristiques qui se rapprochent le plus des caractéristiques latentes obtenues à partir de l'auto-codeur précédent. La figure B7 montre un schéma de la méthode proposée.

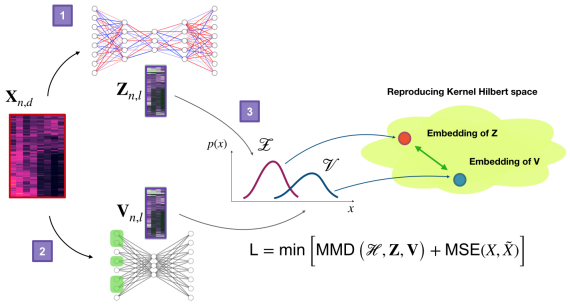

Figure B.7: Taille de la matrice de données $\mathbf{X}$ avant (gauche) et après (droite) la sélection des caractéristiques. d est le nombre de caractéristiques d'entrée et p est le nombre de caractéristiques sélectionnées..

B.5.3 Discussion et conclusion

Ce travail propose les méthodes LKFS et LMMD-FS qui utilisent toutes deux une approche non supervisée pour sélectionner un sous-ensemble de gènes réduit de plus de 8000 gènes à moins de 50 dans trois différents types de jeux de données sur le cancer avec l'objectif d'améliorer la performance du clustering. La sélection des caractéristiques LKFS est effectuée

en améliorant l'alignement entre une matrice de noyau K_z obtenue à partir de l'espace latent $\mathcal{Z}$ appris via un autoencodeur et la matrice résultante K_μ après un apprentissage à noyaux multiples sur des noyaux de caractéristiques. La sélection du LMMD-FS se fait en minimisant la divergence MMD entre l'espace latent de deux auto-codeurs et en construisant un classement des caractéristiques basé sur la norme L1. Les deux approches proposées sont basées sur la construction d'une représentation cible des données d'entrée avec l'espace latent de l'auto-codeur régularisé profond. Cette représentation cible a un niveau de bruit réduit et est de faible dimension. LKFS et LMMD-FS sélectionnent uniquement les caractéristiques qui s'alignent le plus par alignement du noyau et réduisent la distance MMD à la représentation cible respectivement. La représentation cible peut être interprétée comme une distribution préalable qui guide le processus de sélection.
Pour mesurer la qualité du processus de sélection non supervisée des caractéristiques, le score de redondance RED est calculé sur les caractéristiques sélectionnées. En outre, une méthode de clustering K-means est appliquée en aval et la performance du cluster est évaluée par l'homogénéité des étiquettes de vérité du sol à travers les clusters par l'indice Rand.
Les résultats expérimentaux montrent que LKFS surpasse les méthodes LMMD-FS et bechmark par l'indice de Rand et le score RED pour presque tous les ensembles de données et le nombre de caractéristiques sélectionnées. Cela suggère que LKFS est capable de sélectionner des caractéristiques peu redondantes dans un espace d'entrée de grande dimension, ce qui contribue à trouver des clusters bien définis composés principalement d'un sous-type de tumeur. De plus, la méthode LMMD-FS a un comportement similaire à celui des méthodes de référence dans presque tous les jeux de données, quel que soit le nombre de clusters, à l'exception du jeu de données Brain où l'indice Rand obtenu est le deuxième plus élevé.
L'un des avantages de LKFS et LMMD-FS est la représentation latente cible apprise par l'auto-encodeur. Cette représentation capture les caractéristiques saillantes de l'ensemble de données d'entrée et ensuite, par MKL ou MMD, les caractéristiques sélectionnées captureront approximativement la même structure de données. Les méthodes LKFS et LMMD-FS fournissent deux sorties. Le premier est le sous-ensemble de caractéristiques génétiques sélectionnées qui est considérablement réduit par rapport à l'ensemble de caractéristiques d'origine et aide à l'interprétation biologique. Le second est l'espace latent fourni par l'auto-codeur. L'espace latent sert non seulement de représentation cible pour le processus de sélection des caractéristiques, mais aussi d'outil pour l'exploration et l'analyse des données, car il peut résumer dans un espace de dimension inférieure les caractéristiques saillantes des données d'origine.
Une limitation de LKFS et LMMD-FS repose sur la formation de deux modèles, l'auto-codeur régularisé profond et l'auto-codeur MKL ou Vanilla respectivement. Ensuite, les approches de sélection des caractéristiques sont conditionnées par l'architecture et la qualité de l'auto-codeur régularisé profond.

B.6 Fusion multi-omique pour les profils tumoraux

Les données moléculaires issues des profils tumoraux peuvent être divisées en plusieurs groupes de variables aléatoires associées à des mesures effectuées dans des contextes spécifiques. Chacun de ces groupes est une source spécifique de variabilité mesurée dans différents contextes moléculaires. Le scénario multi modal est décrit par des sources de variabilité connues sous le nom de *omique* où chaque *omique* est associé à un type particulier de caractéristiques moléculaires : génome, transcriptome, protéome et métabolome [5]. Chaque omique est composé de macromolécules qui codent des informations biologiques sur la fonction cellulaire en suivant le dogme de la biologie centrale et a un impact différent dans la régulation biologique au sein de la cellule tumorale.
Pour obtenir des données à partir des différents environnements omiques, différentes technologies sont utilisées. Par exemple, la technologie de séquençage de nouvelle génération (NGS) à haut débit mentionnée au chapitre 01 est utilisée pour obtenir la génomique (mutations de l'ADN) et la transcriptomique (concentration d'ARN), tandis que la technologie de spectrométrie de masse [135] est utilisée pour obtenir des données de protéomique (concentration de protéines).
Bien que le coût du séquençage de différentes omiques diminue globalement, la complexité d'obtenir de multiples omiques pour chaque échantillon de cancer est toujours une limitation causée par l'équipement ou la capacité du laboratoire ; pour cette raison, certaines expériences obtiennent des données d'une seule omique, généralement l'*Transcriptomique* et un manque du reste des omiques. Dans cette thèse, du chapitre 02 au chapitre 04, la *transcriptomique* ou l'expression génique *omique* a été utilisée car elle est une source d'information utile pour déterminer l'état d'une cellule donnée et est l'une des omiques les plus utilisées dans l'analyse de cellules uniques et la génomique du cancer [136]. [137] [138].
Néanmoins, si un ensemble de profils tumoraux est décrit par plusieurs omiques au lieu d'apprendre à partir d'un seul, il est possible de proposer des méthodes qui apprennent simultanément à partir de différents omiques afin d'améliorer une tâche d'apprentissage spécifique. Il est supposé que chaque omique fournit des informations spécifiques sur la distribution des données, ainsi l'apprentissage à partir de multiples omiques peut mieux expliquer le manifeste des données.
Les données omiques simples étant de haute dimension, les données multi-omiques le sont encore plus. Ce chapitre étudie et analyse donc s'il est possible d'apprendre un espace latent de faible dimension lorsque plusieurs données omiques sont utilisées comme données d'entrée et de les fusionner en une seule représentation. Pour apprendre un espace latent unique de faible dimension $\mathcal{Z}$ à partir de données multi-omiques, l'approche *Apprentissage multimodal* peut être utilisée en mettant en relation des informations provenant de sources multiples et, par conséquent, améliorer une tâche d'apprentissage en aval dans l'espace latent [139] en supposant que chaque omique peut expliquer une structure particulière des données par elle-même. Dans ce travail, la tâche d'apprentissage en aval est le clustering.

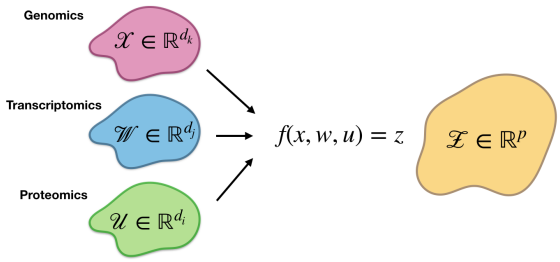

Figure B.8: Schéma du problème de la fusion multi-omique. À partir de multiples omiques $\mathcal{X}$,$\mathcal{W}$,$\mathcal{U}$, une représentation de faible dimension $\mathcal{Z}$ est apprise.

L'approche d'apprentissage multimodal et la fusion de données multi-omiques peuvent être mises en œuvre par différentes stratégies. On peut utiliser des autoencodeurs [139][140] ou des méthodes à noyau, plus précisément une combinaison de noyaux et de kernel-PCA [43].
Une fois l'espace latent appris à partir des données d'entrée multi-omiques, une tâche de sélection de caractéristiques non supervisée peut être effectuée sur les caractéristiques transcriptomiques d'entrée, comme cela a été fait au chapitre 03 avec la méthode de sélection de caractéristiques à noyau latent (LKFS) [129] où l'objectif est de guider, via un espace latent, la sélection d'un sous-ensemble de caractéristiques d'entrée qui conserve la structure de données nécessaire pour trouver des groupes ou des communautés de tumeurs ayant des attributs cliniques communs. La principale différence dans ce travail par rapport à la méthode LKFS est que l'espace latent appris représente une structure de données provenant de multiples omiques au lieu d'une seule. Si l'espace latent appris à partir de données multi-omiques a une meilleure structure qu'un espace latent appris à partir de données mono-omiques sur la base de la performance de regroupement, alors la tâche de sélection de caractéristiques non supervisée de LKFS peut être améliorée si plusieurs omiques sont utilisés pour l'apprentissage de l'espace latent. Dans ce travail, deux données omiques sont utilisées dans la section expérimentale : Mutations somatiques simples (génomique) et RNA-seq (transcriptomique). Pour comprendre l'impact de l'apprentissage de l'espace latent et de la sélection non supervisée des caractéristiques, deux scénarios sont évalués : un seul omique où tous les échantillons sont décrits uniquement par les caractéristiques transcriptomiques et un multi-omique où tous les échantillons sont caractérisés à la fois par la transcriptomique et la génomique.

B.6.1 Fusion multi-omique

Les méthodes de fusion multi-omique visent à apprendre une représentation de faible dimension d'échantillons de tumeurs caractérisés par de multiples données omiques. La représentation de faible dimension capture la variabilité des données omiques d'entrée multiples dans une représentation unique $\mathcal{Z}$. Cela signifie que si la méthode de fusion multi-omique est bien mise en œuvre, elle devrait capturer la structure des données provenant de toutes les données

omiques d'entrée. Ainsi, la fusion multi-omique peut expliquer la structure qu'une seule donnée omique ne peut expliquer à elle seule.
Comme nous l'avons détaillé dans la section consacrée aux travaux connexes, nous étudions dans ce chapitre deux alternatives à la fusion de données multi-omiques. La première est celle des autoencodeurs multimodaux où les données multi-omiques sont utilisées pour apprendre un seul espace latent de faible dimension $\mathcal{Z}$ via une fonction d'encodage et reconstruit de $\mathcal{Z}$ à $\mathcal{X}$ via une fonction de décodage. La seconde alternative utilisée dans ce travail est la combinaison de noyaux, chacun utilisé sur chaque omique d'entrée et ensuite fusionné en un seul noyau dont la représentation correspondante capture la variabilité et la structure de tous les omiques.
Dans ce travail, les deux stratégies sont utilisées pour apprendre une représentation conjointe à partir de données de mutations somatiques simples (génomique) et de séquences d'expression génique (transcriptomique) de profils tumoraux. Les profils tumoraux sont décrits par des variables aléatoires provenant des deux omiques où d_x et d_u sont les dimensions qui définissent les espaces $\mathcal{X} \in \mathbb{R}^{d_x}$ et $\mathcal{U} \in \mathbb{R}^{d_u}$ respectivement. Pour cette raison, un échantillon de tumeur peut être décrit par un seul omic si un seul d'entre eux est disponible ou par les deux dans le cas où les deux sont disponibles. Par conséquent, un ensemble X d'échantillons caractérisés par un seul omic $\mathcal{X}$ est défini comme suit

$$X = \{x_1, \ldots, x_n\}$$

ou caractérisé seulement par $\mathcal{U}$ défini comme

$$U = \{u_1, \ldots, u_n\}$$

et d'autre part, un ensemble d'échantillons multi-omiques est défini lorsque les deux omiques sont disponibles comme suit

$$M_{x,u} = \{m(x_1, u_1), \ldots, m(x_n, u_n)]\}$$

. où n correspond au nombre d'échantillons, x les échantillons décrits uniquement par le premier omic, u les échantillons décrits uniquement par le second omic et m ceux décrits simultanément par les deux omics. Dans la section expérimentale, les méthodes de réduction de la dimensionnalité utilisées sont évaluées sur des ensembles de données mono-omiques et multi-omiques afin d'analyser comment l'apprentissage à partir de données multi-omiques améliore une tâche en aval comme le clustering.
Les sous-sections suivantes expliquent la méthodologie utilisée pour fusionner les données multi-omiques dans une représentation unique $\mathcal{Z}$ par des autoencodeurs et des méthodes à noyau.

B.6.2 Méthode proposée : Sélection de caractéristiques multi-omiques latentes

Le chapitre 03 de cette thèse propose la méthode Latent Kernel Feature Selection (LKFS) conçue pour effectuer une sélection non supervisée de caractéristiques sur des données mono-

omiques. Dans cette section, la méthode LKFS est d'abord brièvement décrite, puis étendue aux données multi-omiques.
La méthode LKFS est conçue pour apprendre tout d'abord un espace latent $z = \phi(x)$ où la cartographie ϕ est un auto-codeur et x sont les données d'apprentissage mono-omiques. Sur l'espace latent appris, un noyau $k_{AE}(z_i, z_j)$ est appris et utilisé pour construire une matrice de noyau K_{AE} qui détaille les mesures de similarité par paire entre les échantillons situés sur $\mathcal{Z}$. Ensuite, à partir d'un ensemble D de matrices noyau de d caractéristiques K_i, une approche d'apprentissage à noyaux multiples effectue une combinaison linéaire $K_\mu = \sum \mu_i K_i$ avec pour objectif de maximiser l'alignement $A(K_{AE}, K_\mu)$ où K_μ est la matrice noyau résultante de la combinaison linéaire et μ est le vecteur solution. Pour maximiser la fonction objectif $A(K_{AE}, K_\mu)$, le vecteur solution peut être clairsemé, ce qui implique que les caractéristiques sélectionnées sont celles pour lesquelles $\mu_i > 0$.
Pour étendre LKFS aux données multi-omiques, un ensemble de données
$M = [m(x_1, u_1), ..., m(x_n, u_n)]$ est utilisé pour entraîner un auto-codeur multimodal au lieu de l'auto-codeur à modalité unique. Un espace latent $\mathcal{Z}$ est obtenu et utilisé comme noyau cible pour guider la tâche de sélection des caractéristiques avec l'étape d'apprentissage à noyaux multiples.

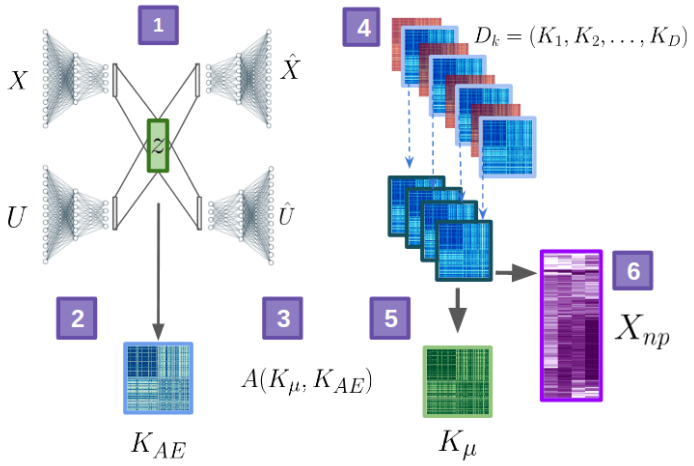

Figure B.9: Proposition d'un pipeline pour la sélection non supervisée de caractéristiques multi-omiques pour le cancer. Tout d'abord, un espace latent est appris et une matrice de noyau cible K_{AE} est construite. Ensuite, un ensemble D de noyaux de caractéristiques est utilisé pour résoudre un problème d'apprentissage à noyaux multiples où le noyau résultant K_μ améliore l'alignement $A(K_{AE}, K_\mu)$. Enfin, un sous-ensemble de caractéristiques p est obtenu en considérant les caractéristiques pour lesquelles le vecteur solution est $\mu_i > 0$.

Les caractéristiques sont toujours sélectionnées à partir d'une seule omique, dans ce cas la transomique. Néanmoins, l'espace latent est appris à partir de la multi-omique. Cela peut faire en sorte que les caractéristiques sélectionnées à partir de la transcriptomique répondent à la structure obtenue à partir de la multi-omique et modifient finalement la solution par

rapport à la tâche de sélection de caractéristiques à partir d'une seule omique.
La figure B9 montre le pipeline de la méthode non supervisée de sélection de caractéristiques multi-omiques proposée.

B.6.3 Résultats et discussion

Ce chapitre explore et définit le problème des données multimodales dans le contexte biomédical du cancer multi-omique. Les données multi-omiques provenant de profils tumoraux sont plus complexes que les données omiques uniques puisque la variabilité provient de différentes sources simultanément. Pour rendre possible l'analyse et l'interprétation des données multi-omiques, des méthodes de réduction de dimension qui fusionnent les différentes modalités en une seule représentation de faible dimension sont nécessaires. Pour gérer ce scénario, des méthodes de réduction de dimension multimodales utilisant des autoencodeurs et des méthodes à noyau sont présentées et décrites.
La conception de la méthode Autoencoder consiste à construire une fonction d'encodage pour chaque modalité omique, puis ces fonctions sont fusionnées en aval dans le réseau de neurones feedforward en une seule représentation de faible dimension z_{multi} connue sous le nom d'espace latent et détaillée dans la figure 2. Pour reconstruire les données d'entrée à partir de l'espace latent, deux fonctions de décodage sont séparées. L'auto-codeur multimodal tente de reconstruire les données d'entrée multi-omiques ; ainsi, la fonction de perte utilisée pour apprendre les paramètres du réseau comporte un terme pour chaque omique. En outre, à des fins expérimentales, des auto-codeurs standard à modalité unique sont utilisés sur chaque ensemble de données omiques uniques afin de comparer si la fusion de multiples omiques améliore la qualité de l'espace latent résultant.
Une autre alternative proposée pour combiner et fusionner des données multimodales est la méthode de combinaison de noyaux. Elle consiste à apprendre un noyau k_{multi} à partir de la combinaison des noyaux k_{ssm} et k_{exp} correspondant respectivement aux modalités génomiques et transcriptomiques. Ensuite, en utilisant les k_{multi} dans un algorithme d'ACP à noyau, il est possible d'apprendre l'espace latent de faible dimension z_{multi} du domaine multi-omique.
Comme expliqué dans la section matériaux et méthodes, l'évaluation de l'espace latent résultant est faite par une tâche aval non supervisée, plus précisément une tâche de clustering par la méthode des k-means. La qualité de l'espace latent résultant est évaluée par la façon dont les profils de tumeurs sont regroupés après l'étape de réduction de dimension selon les étiquettes de sous-type de tumeur. Avec les sous-types de tumeurs comme étiquettes de vérité fondamentale, un espace latent qui tend à regrouper les profils de tumeurs du même sous-type est préféré à ceux qui tendent à distribuer différents sous-types de tumeurs ensemble.
L'espace latent multi-omique qui en résulte peut être utilisé pour étendre la méthode LKFS [129] présentée dans le chapitre 03 aux omiques multiples et améliorer la performance de la sélection des caractéristiques. Une tâche de sélection de caractéristiques multi-omiques est présentée et évaluée.
Les résultats expérimentaux sont obtenus en effectuant une réduction de la dimensionnalité avec chaque méthode proposée en utilisant des ensembles de données mono-omiques et

multi-omiques. Pour évaluer l'espace latent, la performance de clustering est mesurée et on observe que la meilleure performance est obtenue sur l'espace latent appris par l'autoencodeur multi-modal. Les résultats obtenus à l'aide de l'auto-codeur mono-omique sont inférieurs à ceux obtenus à l'aide de l'auto-codeur multimodal formé sur les ensembles de données multi-omiques. Les résultats obtenus par la méthode de combinaison de noyaux sont inférieurs à ceux obtenus par l'auto-codeur. De plus, on observe que la méthode de combinaison de noyaux ne montre aucune amélioration entre l'espace latent obtenu par des données transcriptomiques d'expression génique unique et le scénario multi-omique, ce qui suggère que la méthode de combinaison de noyaux ne peut pas capturer les informations de la modalité génomique. Les travaux futurs devraient envisager d'autres types de combinaison de noyaux, tels que $K_{multi}((x, u), (x^{'}, u^{'}))$.

Puisque la méthode d'autoencodage multimodal surpasse la méthode de combinaison à noyau pour le problème de fusion de données multi-omique, ces résultats suggèrent que l'approche autoencodeur améliore la qualité de l'espace latent résultant par rapport aux méthodes à noyau avec un coût de surparamétrage. L'amélioration du regroupement dans l'espace latent a un impact sur la qualité de la tâche de sélection des caractéristiques, montrant que les caractéristiques sélectionnées via un espace latent multi-omique surpassent celles sélectionnées via un espace latent mono-omique.

Dans ce chapitre, deux approches ont été présentées pour traiter le problème de la fusion de données multimodales à l'aide de données synthétiques et réelles de panceras et une méthode de sélection de caractéristiques multi-omiques est présentée pour sélectionner des caractéristiques basées sur un espace latent multi-omique. Comme les dépôts de données multi-omiques sur le cancer ne cessent de croître, la nécessité d'étudier et de développer des approches multimodales sera une priorité en génomique du cancer.

Bibliography

[1] P.-Y. Wu, C.-W. Cheng, C. D. Kaddi, J. Venugopalan, R. Hoffman, and M. D. Wang, IEEE Transactions on Biomedical Engineering **64**, 263 (2016).

[2] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*, vol. 1 (MIT press Cambridge, 2016).

[3] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart, Nature genetics **45**, 1113 (2013).

[4] M. R. Stratton, P. J. Campbell, and P. A. Futreal, Nature **458**, 719 (2009).

[5] D. Ramazzotti, A. Lal, B. Wang, S. Batzoglou, and A. Sidow, Nature communications **9**, 1 (2018).

[6] D. Lahat, T. Adali, and C. Jutten, Proceedings of the IEEE **103**, 1449 (2015).

[7] B. A. Ponder, Nature **411**, 336 (2001).

[8] H. Moch, T. Gasser, M. B. Amin, J. Torhorst, G. Sauter, and M. J. Mihatsch, Cancer **89**, 604 (2000).

[9] S. C. Schuster, Nature methods **5**, 16 (2008).

[10] Z. Wang, M. Gerstein, and M. Snyder, Nature reviews genetics **10**, 57 (2009).

[11] P. Liang and A. B. Pardee, Nature Reviews Cancer **3**, 869 (2003).

[12] S. Oota, Methods (2019).

[13] R. Sager, Proceedings of the National Academy of Sciences **94**, 952 (1997).

[14] C. L. Nutt, D. Mani, R. A. Betensky, P. Tamayo, J. G. Cairncross, C. Ladd, U. Pohl, C. Hartmann, M. E. McLaughlin, T. T. Batchelor, et al., Cancer research **63**, 1602 (2003).

[15] L. J. Van't Veer, H. Dai, M. J. Van De Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. Van Der Kooy, M. J. Marton, A. T. Witteveen, et al., nature **415**, 530 (2002).

[16] I. C. G. Consortium et al., Nature **464**, 993 (2010).

[17] J. Lapointe, C. Li, J. P. Higgins, M. Van De Rijn, E. Bair, K. Montgomery, M. Ferrari, L. Egevad, W. Rayford, U. Bergerheim, et al., Proceedings of the National Academy of Sciences **101**, 811 (2004).

[18] J. Shawe-Taylor, N. Cristianini, et al., *Kernel methods for pattern analysis* (Cambridge university press, 2004).

[19] G. V. Trunk, IEEE Transactions on pattern analysis and machine intelligence pp. 306–307 (1979).

[20] L. Amsaleg, O. Chelly, T. Furon, S. Girard, M. E. Houle, K.-i. Kawarabayashi, and M. Nett, in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015), pp. 29–38.

[21] Y. Lu and J. Han, Information Systems **28**, 243 (2003).

[22] I. Guyon and A. Elisseeff, Journal of machine learning research **3**, 1157 (2003).

[23] A.-C. Haury, P. Gestraud, and J.-P. Vert, PloS one **6** (2011).

[24] Z. He and W. Yu, Computational biology and chemistry **34**, 215 (2010).

[25] N. Almugren and H. Alshamlan, IEEE Access **7**, 78533 (2019).

[26] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed, IEEE/ACM transactions on computational biology and bioinformatics **13**, 971 (2015).

[27] Y. Saeys, I. Inza, and P. Larrañaga, bioinformatics **23**, 2507 (2007).

[28] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf, arXiv preprint arXiv:1605.09522 (2016).

[29] B. Schölkopf, K. Tsuda, and J. Vert, MIT Press, Cambridge Scsibrany H, Karlovits M, Demuth W, Müller F, Varmuza K (2003) Clustering and similarity of chemical structures represented by binary substructure descriptors. Chemometr Intell Lab Syst **67**, 95 (2004).

[30] J.-P. Vert, in *Kernel methods in bioengineering, signal and image processing* (IGI Global, 2007), pp. 42–63.

[31] M. Palazzo, P. Yankilevich, and P. Beauseroy, arXiv preprint arXiv:2004.04866 (2020).

[32] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, vol. 1 (Springer series in statistics New York, 2001).

[33] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, vol. 112 (Springer, 2013).

[34] B. E. Boser, I. M. Guyon, and V. N. Vapnik, in *Proceedings of the fifth annual workshop on Computational learning theory* (1992), pp. 144–152.

[35] C. Cortes and V. Vapnik, Machine learning **20**, 273 (1995).

[36] F. Melo, Encyclopedia of Systems Biology pp. 38–39 (2013).

[37] J.-P. Vert, K. Tsuda, and B. Schölkopf, Kernel methods in computational biology **47**, 35 (2004).

[38] J. Kandola, J. Shawe-Taylor, and N. Cristianini (2002).

[39] N. Cristianini, J. Kandola, A. Elisseeff, and J. Shawe-Taylor, in *Innovations in Machine Learning* (Springer, 2006), pp. 205–256.

[40] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. S. Kandola, in *Advances in neural information processing systems* (2002), pp. 367–373.

[41] J. Kandola, J. Shawe-Taylor, and N. Cristianini (2002).

[42] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf, The Journal of Machine Learning Research **7**, 1531 (2006).

[43] M. Gönen and E. Alpaydın, Journal of machine learning research **12**, 2211 (2011).

[44] J.-B. Pothin and C. Richard, in *2006 14th European Signal Processing Conference* (IEEE, 2006), pp. 1–4.

[45] C. Cortes, M. Mohri, and A. Rostamizadeh, Journal of Machine Learning Research **13**, 795 (2012).

[46] S. Zhong, D. Chen, Q. Xu, and T. Chen, Pattern Recognition **46**, 2045 (2013).

[47] K.-B. Duan, J. C. Rajapakse, H. Wang, and F. Azuaje, IEEE transactions on nanobioscience **4**, 228 (2005).

[48] M. Yamada, W. Jitkrittum, L. Sigal, E. P. Xing, and M. Sugiyama, Neural computation **26**, 185 (2014).

[49] Y. Zhang, A. Li, C. Peng, and M. Wang, IEEE/ACM transactions on computational biology and bioinformatics **13**, 825 (2016).

[50] A. Adorada, R. Permatasari, P. W. Wirawan, A. Wibowo, and A. Sujiwo, in *2018 2nd International Conference on Informatics and Computational Sciences (ICICoS)* (IEEE, 2018), pp. 1–4.

[51] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, Machine learning **46**, 389 (2002).

[52] X. Huang, L. Zhang, B. Wang, F. Li, and Z. Zhang, Applied Intelligence **48**, 594 (2018).

[53] H. Peng, F. Long, and C. Ding, IEEE Transactions on Pattern Analysis & Machine Intelligence pp. 1226–1238 (2005).

[54] M. Radovic, M. Ghalwash, N. Filipovic, and Z. Obradovic, BMC bioinformatics **18**, 1 (2017).

[55] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, in *International conference on algorithmic learning theory* (Springer, 2005), pp. 63–77.

[56] Z. Zhao, L. Wang, and H. Liu, in *Twenty-fourth AAAI conference on artificial intelligence* (2010).

[57] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, ACM Computing Surveys (CSUR) **50**, 1 (2017).

[58] J. Reunanen, Journal of Machine Learning Research **3**, 1371 (2003).

[59] S. Mika, B. Schölkopf, A. J. Smola, K.-R. Müller, M. Scholz, and G. Rätsch, in *Advances in neural information processing systems* (1999), pp. 536–542.

[60] C. M. Bishop, *Pattern recognition and machine learning* (springer, 2006).

[61] B. Schölkopf, in *Advances in neural information processing systems* (2001), pp. 301–307.

[62] B. Schölkopf, A. J. Smola, F. Bach, et al., *Learning with kernels: support vector machines, regularization, optimization, and beyond* (MIT press, 2002).

[63] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Journal of Machine Learning Research **12**, 2825 (2011).

[64] A. Spira, J. E. Beane, V. Shah, K. Steiling, G. Liu, F. Schembri, S. Gilman, Y.-M. Dumas, P. Calner, P. Sebastiani, et al., Nature medicine **13**, 361 (2007).

[65] R. Edgar, M. Domrachev, and A. E. Lash, Nucleic acids research **30**, 207 (2002).

[66] L. Chen, H. M. Linden, B. O. Anderson, and C. I. Li, Breast cancer research and treatment **147**, 609 (2014).

[67] M. Palazzo, P. Beauseroy, and P. Yankilevich, BMC bioinformatics **20**, 655 (2019).

[68] D. Kobak and P. Berens, Nature communications **10**, 1 (2019).

[69] M. Palazzo, P. Beauseroy, D. Koile, and P. Yankilevich, in *IV Simposio Argentino de GRANdes DAtos (AGRANDA 2018)-JAIIO 47 (CABA, 2018)* (2018).

[70] M. Manzi, M. Palazzo, M. E. Knott, P. Beauseroy, P. Yankilevich, M. I. Giménez, and M. E. Monge, Journal of Proteome Research **20**, 841 (2020).

[71] M. Palazzo, P. Beauseroy, and P. Yankilevich, Electronic Journal of SADIO (EJS) **18**, 26 (2019).

[72] M. Palazzo, P. Yankilevich, and P. Beauseroy (????).

[73] S. Wold, K. Esbensen, and P. Geladi, Chemometrics and intelligent laboratory systems **2**, 37 (1987).

[74] S. Mika, B. Schölkopf, A. J. Smola, K.-R. Müller, M. Scholz, and G. Rätsch, in *NIPS* (1998), vol. 11, pp. 536–542.

[75] G. E. Hinton and R. R. Salakhutdinov, science **313**, 504 (2006).

[76] L. Van der Maaten and G. Hinton, Journal of machine learning research **9** (2008).

[77] K.-C. Li, Journal of the American Statistical Association **86**, 316 (1991).

[78] T. Suzuki and M. Sugiyama, in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (2010), pp. 804–811.

[79] Y. Dong, Journal of Statistical Planning and Inference (2020).

[80] L. Li and H. Li, Bioinformatics **20**, 3406 (2004).

[81] H.-M. Hsueh and C.-A. Tsai, BMC bioinformatics **17**, 74 (2016).

[82] K. Fukumizu, F. R. Bach, and M. I. Jordan (2006).

[83] K. P. Adragni and R. D. Cook, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **367**, 4385 (2009).

[84] A. Globerson and N. Tishby, Journal of Machine Learning Research **3**, 1307 (2003).

[85] X. Yin and H. Hilafu, Journal of the Royal Statistical Society: Series B: Statistical Methodology pp. 879–892 (2015).

[86] A. Berlinet and C. Thomas-Agnan, *Reproducing kernel Hilbert spaces in probability and statistics* (Springer Science & Business Media, 2011).

[87] K. Fukumizu, F. R. Bach, M. I. Jordan, et al., The Annals of Statistics **37**, 1871 (2009).

[88] M. Wang, F. Sha, and M. I. Jordan, in *Advances in Neural Information Processing Systems* (2010), pp. 2379–2387.

[89] L. McInnes, J. Healy, and J. Melville, arXiv preprint arXiv:1802.03426 (2018).

[90] J. Zhang, R. Bajari, D. Andric, F. Gerthoffert, A. Lepsa, H. Nahal-Bose, L. D. Stein, and V. Ferretti, Nature biotechnology **37**, 367 (2019).

[91] D. R. Cox, Journal of the Royal Statistical Society: Series B (Methodological) **34**, 187 (1972).

[92] S. Pölsterl, Journal of Machine Learning Research **21**, 1 (2020), URL `http://jmlr.org/papers/v21/20-729.html`.

[93] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati, Jama **247**, 2543 (1982).

[94] H. Uno, T. Cai, M. J. Pencina, R. B. D'Agostino, and L.-J. Wei, Statistics in medicine **30**, 1105 (2011).

[95] S. Pölsterl, N. Navab, and A. Katouzian, in *Machine Learning and Knowledge Discovery in Databases*, edited by A. Appice, P. P. Rodrigues, V. Santos Costa, J. Gama, A. Jorge, and C. Soares (2015), Lecture Notes in Computer Science, pp. 243–259.

[96] L. Jiang, Y. Xiao, Y. Ding, J. Tang, and F. Guo, Frontiers in genetics **10**, 20 (2019).

[97] R. Chen, M. Smith-Cohn, A. L. Cohen, and H. Colman, Neurotherapeutics **14**, 284 (2017).

[98] J. Xiao, X. Lu, X. Chen, Y. Zou, A. Liu, W. Li, B. He, S. He, and Q. Chen, Oncotarget **8**, 71759 (2017).

[99] F. Chen, Y. Zhang, Y. Şenbabaoğlu, G. Ciriello, L. Yang, E. Reznik, B. Shuch, G. Micevic, G. De Velasco, E. Shinbrot, et al., Cell reports **14**, 2476 (2016).

[100] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, ACM Computing Surveys (CSUR) **50**, 94 (2018).

[101] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, The Journal of Machine Learning Research **13**, 723 (2012).

[102] J. G. Dy and C. E. Brodley, Journal of machine learning research **5**, 845 (2004).

[103] A. K. Jain, Pattern recognition letters **31**, 651 (2010).

[104] X. Zheng, Q. Lei, R. Yao, Y. Gong, and Q. Yin, EURASIP Journal on Image and Video Processing **2018**, 68 (2018).

[105] I. S. Dhillon, Y. Guan, and B. Kulis, in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (2004), pp. 551–556.

[106] K. Chaudhary, O. B. Poirion, L. Lu, and L. X. Garmire, Clinical Cancer Research **24**, 1248 (2018).

[107] C. Elkan, in *Proceedings of the 20th international conference on Machine Learning (ICML-03)* (2003), pp. 147–153.

[108] W. M. Rand, Journal of the American Statistical association **66**, 846 (1971).

[109] Z. M. Hira and D. F. Gillies, Advances in bioinformatics **2015** (2015).

[110] Y. Guo, X. Shang, and Z. Li, Neurocomputing **324**, 20 (2019).

[111] G. P. Way and C. S. Greene, BioRxiv p. 174474 (2017).

[112] Z. Wang and Y. Wang, in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (IEEE, 2018), pp. 1286–1289.

[113] C. H. Grønbech, M. F. Vording, P. N. Timshel, C. K. Sønderby, T. H. Pers, and O. Winther, bioRxiv p. 318295 (2018).

[114] M. Kampffmeyer, S. Løkse, F. M. Bianchi, R. Jenssen, and L. Livi, in *Scandinavian Conference on Image Analysis* (Springer, 2017), pp. 419–430.

[115] D. P. Kingma and J. Ba, arXiv preprint arXiv:1412.6980 (2014).

[116] S. Ioffe and C. Szegedy, arXiv preprint arXiv:1502.03167 (2015).

[117] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, Journal of Machine Learning Research **9**, 2491 (2008).

[118] W. Du, Z. Cao, T. Song, Y. Li, and Y. Liang, BioData mining **10**, 4 (2017).

[119] C. Ding and H. Peng, Journal of bioinformatics and computational biology **3**, 185 (2005).

[120] M. Moon and K. Nakai, BMC genomics **17**, 1026 (2016).

[121] H. Zou and T. Hastie, Journal of the royal statistical society: series B (statistical methodology) **67**, 301 (2005).

[122] M. B. Lopes, A. Veríssimo, E. Carrasquinha, S. Casimiro, N. Beerenwinkel, and S. Vinga, BMC bioinformatics **19**, 168 (2018).

[123] S. Alelyani, J. Tang, and H. Liu, in *Data Clustering* (Chapman and Hall/CRC, 2018), pp. 29–60.

[124] D. Cai, C. Zhang, and X. He, in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (ACM, 2010), pp. 333–342.

[125] S. Bandyopadhyay, D. Ghosh, R. Mitra, and Z. Zhao, Scientific reports **5**, 8004 (2015).

[126] D. M. Witten and R. Tibshirani, Journal of the American Statistical Association **105**, 713 (2010).

[127] Z. Zhao and H. Liu, in *Proceedings of the 24th international conference on Machine learning* (ACM, 2007), pp. 1151–1157.

[128] Z. Zhao, J. Wang, S. Sharma, N. Agarwal, H. Liu, and Y. Chang, in *Proceedings of the 2010 SIAM International Conference on Data Mining* (SIAM, 2010), pp. 838–849.

[129] M. Palazzo, P. Beauseroy, and P. Yankilevich, arXiv preprint arXiv:2007.06106 (2020).

[130] A. Mirzaei, V. Pourahmadi, M. Soltani, and H. Sheikhzadeh, Neurocomputing **383**, 396 (2020).

[131] G. K. Dziugaite, D. M. Roy, and Z. Ghahramani, arXiv preprint arXiv:1505.03906 (2015).

[132] S. Zhao, J. Song, and S. Ermon, arXiv preprint arXiv:1706.02262 (2017).

[133] L. v. d. Maaten and G. Hinton, Journal of machine learning research **9**, 2579 (2008).

[134] R. A. van den Berg, H. C. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf, BMC genomics **7**, 142 (2006).

[135] R. Aebersold and M. Mann, Nature **422**, 198 (2003).

[136] Z. Huang, T. S. Johnson, Z. Han, B. Helm, S. Cao, C. Zhang, P. Salama, M. Rizkalla, C. Y. Yu, J. Cheng, et al., BMC medical genomics **13**, 1 (2020).

[137] S. Cascianelli, I. Molineris, C. Isella, M. Masseroli, and E. Medico, Scientific reports **10**, 1 (2020).

[138] C. Huang, E. A. Clayton, L. V. Matyunina, L. D. McDonald, B. B. Benigno, F. Vannberg, and J. F. McDonald, Scientific reports **8**, 1 (2018).

[139] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, in *ICML* (2011).

[140] K. Tan, W. Huang, J. Hu, and S. Dong, BMC Medical Informatics and Decision Making **20**, 1 (2020).

[141] B. Zhu, N. Song, R. Shen, A. Arora, M. J. Machiela, L. Song, M. T. Landi, D. Ghosh, N. Chatterjee, V. Baladandayuthapani, et al., Scientific reports **7**, 16954 (2017).

[142] G. P. Way and C. S. Greene, arXiv preprint arXiv:1711.04828 (2017).

[143] H.-I. H. Chen, Y.-C. Chiu, T. Zhang, S. Zhang, Y. Huang, and Y. Chen, BMC systems biology **12**, 45 (2018).

[144] J. Iranzo, I. Martincorena, and E. V. Koonin, Proceedings of the National Academy of Sciences **115**, E6010 (2018).

[145] S. Kim, H. Lee, K. Kim, and J. Kang, BMC medical genomics **11**, 33 (2018).

[146] J. Mariette and N. Villa-Vialaneix, Bioinformatics **34**, 1009 (2018).

[147] M. Tao, T. Song, W. Du, S. Han, C. Zuo, Y. Li, Y. Wang, and Z. Yang, Genes **10**, 200 (2019).

Martin PALAZZO

Doctorat : Optimisation et Sureté des Systèmes

Année 2021

Réduction de la dimensionnalité des profils tumoraux biomédicaux : une approche d'apprentissage automatique

Le rythme croissant de génération de données à partir de profils tumoraux au cours de la dernière décennie a permis le développement d'algorithmes d'apprentissage statistique pour explorer et analyser le paysage des types et sous-types de tumeurs et la survie des patients d'un point de vue biomoléculaire. Les données tumorales sont principalement décrites par des caractéristiques transcriptomiques et le niveau d'expression d'un transcrit génique donné dans la cellule tumorale. Par conséquent, ces caractéristiques peuvent être utilisées pour apprendre des règles statistiques qui améliorent la compréhension de l'état et du type d'une cellule cancéreuse.
Néanmoins, les données tumorales transcriptomiques sont de grande dimension et chaque tumeur peut être décrite par des milliers de caractéristiques génétiques, ce qui rend difficile la réalisation d'une tâche d'apprentissage automatique et la compréhension des mécanismes biologiques sous-jacents. Cette thèse étudie comment réduire la dimensionnalité et gagner en interprétabilité pour savoir quels gènes codent le signal de la distribution des données en proposant des méthodes de réduction de dimension basées sur un modèle qui envisage la structure globale des données à l'aide d'un espace de représentation latente. Les méthodes proposées ont montré des améliorations dans les tâches de sélection de caractéristiques supervisées et non supervisées par rapport aux méthodes de référence pour classer et apprendre des sous-groupes de tumeurs respectivement.

Mots clés : apprentissage automatique – réseaux neuronaux (informatique) – noyaux (analyse fonctionnelle) – génomique – cancer – réduction de dimension.

Dimensionality Reduction of Biomedical Tumor Profiles: a Machine Learning Approach

The increasing pace of data generation from tumor profiles during the last decade has enable the development of statistical learning algorithms to explore and analyze the landscape of tumor types, subtypes and patient survival from a biomolecular point of view. Tumor data is mainly described by trasncriptomic features and the level of expression of a given gene-transcript in the tumor cell, therefore these features can be used to learn statistical rules that improves the understanding about the state and type of a cancer cell.
Nevertheless transcriptomic tumor data is high dimensional and each tumor can be described by thousands of gene features making it difficult to perform a machine learning task and to understand the underlying biological mechanisms. This thesis studies how to reduce dimensionality and to gain interpretability about which genes encode signals of the data distribution by proposing dimension reduction methods based on Feature Selection and Feature Extraction pipelines. The proposed methods are based on Latent Variable Models and Kernel Methods with the idea to explore the connection between pair-wise similarity functions of tumor samples and low dimensional latent spaces that captures the inner structure of the training data. Proposed methods have shown improvements in supervised and unsupervised feature selection tasks when compared with benchmark methods to classify and learn subgroups of tumors respectively.

Keywords: machine learning – neural network (computer science) – Kernel functions – genomics – cancer – dimensionality reduction.

Thèse réalisée en partenariat entre :

UTN.BA
UNIVERSIDAD TECNOLÓGICA NACIONAL
FACULTAD REGIONAL BUENOS AIRES

utt
UNIVERSITÉ DE TECHNOLOGIE
TROYES

Ecole Doctorale "Sciences pour l'Ingénieur"