UNIVERSIDAD
DE GRANADA

Departamento de Arquitectura y Tecnología de Computadores

TESIS DOCTORAL

Development of advanced machine learning models for the fusion of heterogeneous biological sources in clinical decision support systems for cancer

Autor:
Francisco Carrillo Pérez
franciscocp@ugr.es

Directores:
Luis Javier Herrera Maldonado
Ignacio Rojas Ruiz

Granada, noviembre de 2022

Editor: Universidad de Granada. Tesis Doctorales
Autor: Francisco Carrillo Pérez
ISBN: 978-84-1117-668-2
URI: https://hdl.handle.net/10481/79672

UNIVERSIDAD
DE GRANADA

Development of advanced machine learning models for the fusion of heterogeneous biological sources in clinical decision support systems for cancer

Tesis para la obtención del título de doctor por la Universidad de Granada
Programa de doctorado en Tecnologías de la Información y la Comunicación

Francisco Carrillo Pérez
franciscocp@ugr.es

Directores de tesis

Dr. D. Luis Javier Herrera Maldonado

Circuitos y Sistemas para Procesamiento de la Información (Universidad de Granada)

Dr. D. Ignacio Rojas Ruiz

Circuitos y Sistemas para Procesamiento de la Información (Universidad de Granada)

Development of advanced machine learning models for the fusion of heterogeneous biological sources in clinical decision support systems for cancer
Francisco Carrillo Pérez

Funding
This doctoral thesis was funded by the predoctoral program Contratos Predoctorales para la Formación de Doctores (ref. RTI2018-101674-B-I00) from the Spanish Ministry of Universities.

Colophon
This document was typeset with the help of KOMA-Script and LaTeX using the kaobook class.

Acknowledgments

There are a lot of people I need to be thankful for this thesis. We are a compound of all the people we learn from and that help us every day. Therefore, I am what I have learnt from others.

Thanks to Luis Javier and Ignacio, my thesis supervisors, who have been incredibly supportive during this process. I have learnt from their experience, and with kindness, they have improved my weaknesses. I am a better researcher today thanks to them.

Thanks to my parents and family for their support during this process. Without their love and the opportunities provided by them, I wouldn't have been as ready as I was for facing the research environment, and I am really thankful for that. They indirectly taught me, based on their actions, that you need to work hard and always give your maximum while staying humble. It doesn't matter to stay in the background if you are proud of your work, and that is something that defines how I approach life.

Thanks to my mentors, colleagues and all the people I have worked with. I have been extremely lucky to work with amazing researchers I have learnt from. Thanks to Alberto for all his support and lessons during my undergrad and master's. Thanks to Olivier for his incredible support and teachings. I have greatly grown as a researcher during my time and Stanford, and I need to be thankful for all the opportunities that have brought me, and for his constant support of my "crazy" ideas. Thanks to Marija for all the laughs and the unforgivable number of hours spent together in those offices without windows. You are an amazing researcher, and you will have a future brighter than mine. Thanks to Juan Carlos for his support and laughs during the first period of this thesis. And of course, thanks to all my friends for their support during this process and the different steps I have made. Those steps would have not been possible without you.

Thanks to Yésica. I cannot imagine a better partner in crime for life than you, and even though we have spent many months apart, I have never felt lonely. Thanks for your support during the thesis, thanks for teaching me so much about biology and life, thanks for making me laugh until my belly hurts, and thanks for loving me. We have already lived together in three different countries, so let's see if we can increase the list. If I can be a support in your thesis just a fraction of how much you have been to me I would be more than happy. I love you.

Finally, thanks to all the researchers worldwide, and those before me. As Newton said, "If I have seen further, it is because I have stood on the shoulders of giants", and so do I (not comparing my poor self with Newton).

Abstract

Development of advanced machine learning models for the fusion of heterogeneous biological sources in clinical decision support systems for cancer

Cancer is one of the leading causes of death worldwide, just behind cardiovascular diseases. An early diagnosis is key for the prognosis of the patient, since it allows applying the most suitable treatment. To do so, multiple screenings are routinely performed on the patient involving, for instance, the visual examination of histopathological slides, the analysis of the clinical history, or finding alterations in their gene expression. These examinations, however, are usually time-consuming, and not always the physicians have the experience to analyze them. To help them with these tasks, clinical decision support systems have been created in recent years using the advances in the machine learning field. Machine learning models are able to automatically learn from these data, and find insights that can help them to solve a specific task. This is part of the precision medicine field where, using a data-driven approach, we tailor the diagnosis, treatment, and other clinical outcomes to the specific characteristics of the patient. Thanks to the advances in this field, more heterogeneous sources of biological information are being gathered, and they provide diverse features that can help to accurately diagnose a cancer patient. This allows to create systems that use all the available information, accurately modelling the patient's disease. This would be similar to having a separate diagnosis per data modality from a group of expert clinicians, where the final diagnosis is based on their analysis of their source of expertise. Unfortunately, not all these sources are always available, limiting the potential of creating multi-modal machine learning models.

In this thesis, we explore the improvements that can be obtained by using multi-modal machine learning models resilient to missing modalities over single-modality ones in the area of cancer diagnosis. Firstly, we tackled the problem of lung cancer subtyping diagnosis using two of the most-used biomedical modalities in literature (gene expression and histopathology images), showing the improvements that can be obtained by fusing these two modalities in comparison to being independently used. Next, to study the limits that can be achieved by fusing heterogeneous biological sources, we include three new modalities to the proposed problem (micro-RNA, DNA Methylation values, and the copy number variation of the genes). We tested which modalities complemented each other, and which is the performance that can be obtained by fusing all these modalities in a classification model. Lastly, we approached the problem of data scarcity in biomedical multi-modal problems, presenting advance methodologies for biological data generation. Inspired by the recent advances in multi-modal generative models for natural images, we focus on generating one modality based on a paired one (RNA-to-image synthesis problem) for healthy tissues. We showed how the synthetic generated data were similar to the real samples and the model was able to impute missing modalities.

Resumen

Diseño de modelos avanzados de aprendizaje maquina para la integracion de fuentes biologicas heterogeneas en sistemas de ayuda al diagnostico del cancer

El cáncer es una de las primeras causas de mortalidad en el mundo, solo por detrás de las enfermedades cardiovasculares. Poder realizar un diagnóstico temprano es crucial para mejorar la esperanza de vida del paciente, ya que se le podría proporcionar un tratamiento más eficaz y adecuado a su estado. Para poder realizar este diagnóstico, múltiples pruebas médicas se le realizan rutinariamente a un paciente. Entre ellas, se incluye la inspección visual de imágenes histológicas, el análisis de la historia clínica, o encontrar alteraciones en la expresión de gen del paciente. Sin embargo, estas pruebas conllevan bastante tiempo, y no todos los hospitales están equipados con el material necesario para su realización. Con el fin de ayudar a los médicos en estas tareas de análisis, y gracias a los avances en el campo del aprendizaje automático, se han ido creado sistemas de apoyo al diagnóstico en los últimos años. Los algoritmos de aprendizaje máquina son capaces de aprender automáticamente de estos datos, y encontrar patrones que les ayuden a resolver una tarea específica. Esto forma parte del área de la medicina de precisión en la que, siguiendo una metodología basada en datos, se puede ofrecer un diagnóstico más robusto o elegir un tratamiento más eficaz basado en las características genéticas o del historial médico del paciente entre otras. Gracias en parte a los avances en este área, cada vez se recogen más fuentes de información biológica heterogénea, las cuáles proporcionan importante información biológica que pueden ayudar a la hora de realizar el diagnóstico de un paciente. Esto abre la posibilidad de crear sistemas que utilicen toda esta información, describiendo mejor la patología del paciente. Esto es similar a tener en cuenta la opinión de distintos especialistas a la hora de realizar un diagnóstico, donde cada uno de ellos se basa en una fuente de datos distinta. Desafortunadamente, no todas las fuentes de información están siempre disponibles, lo que limita la creación de algoritmos de aprendizaje máquina multimodales.

En esta tesis, exploramos las mejoras que se pueden obtener haciendo uso de algoritmos de aprendizaje máquina multimodales en comparación con aquellos que utilizan una única modalidad. En primer lugar, hacemos uso de los dos tipos de datos más usados en la literatura (expresión de gen e imágenes histológicas) para el diagnóstico de los distintos subtipos de cáncer de pulmón, mostrando las mejoras que se pueden obtener haciendo uso de estas dos modalidades en conjunto en lugar de por separado. A continuación, para estudiar los límites que se pueden alcanzar integrando fuentes biológicas heterogéneas, añadimos tres modalidades adicionales (micro-RNA, datos de metilación del ADN, e información de la variación en el número de copias de los genes) para el mismo problema. Comprobamos qué modalidades interaccionan mejor con cuáles, y cuál es el límite que se puede alcanzar al integrar todas estas modalidades en un único modelo de clasificación. Por último, afrontamos el problema de la escasez de datos en problemas biomédicos multimodales aportando metodologías avanzadas de generación de datos sintéticos biológicos. Inspirados por los recientes avances en modelos generativos multimodales para imágenes no biológicas, nos enfocamos en la generación de una modalidad basándonos en su par (el problema de la síntesis de imagen histológicas en base a la expresión de gen), para tejidos sanos. Demostramos como los datos sintéticos generados se asemejan a los datos reales y pueden servir para la imputación de modalidades faltantes.

Contents

List of Figures

List of Tables

Fundaments

Fundamentals of cancer biology 1

1.1 Introduction

Biology is the study of life. The word "biology" is derived from the Greek word "bios" (life) and "logos" (study). In general, biologists study the structure, function, growth, origin, evolution, and distribution of living organisms. Specifically, human biology is an area of research focused on studying humans through the interactions between many diverse fields (such as genetics, evolution, physiology, anatomy, epidemiology, or population genetics) [1]. With the recent advances in information technology, the quantity of human biological data collected has not stopped increasing. That ranges from clinical to single-cell expression data, passing through other omics (more than 25, and the number is growing [2, 3]: genomics, epigenomics, microbiomics, lipidomics, proteomics, glycomics, foodomics, transcriptomics, and metabolomics just to mention a few) allowing to study humans in a multi-scale way. These advances have especially affected the study of human diseases, allowing us to model them from different levels, and increasing the generated knowledge.

Within human diseases, cancer is one of the leading causes of death worldwide. It accounted for nearly 10 million deaths in 2020, and it is the second deadliest disease worldwide just after cardiovascular diseases [4]. An early diagnosis of the disease is crucial for a good prognosis for the patient, and therefore, multiple screenings are usually carried out in clinical practice. These data are becoming increasingly varied, ranging from patient-level clinical variables (e.g. age, smoking history) to sequencing the genome of the patient to find changes in the expression of the biological biomarkers or digitized cancer tissue slides [5]. This allows to study the disease in a multi-scale and multi-omic approach, levering the information from all the available sources [6]. However, manually inspecting the collected data is unfeasible, given their huge size. Thus, new solutions arise for dealing with these data, based on high-performance computing (HPC). The combination of big quantities of data and high-performance computing offers new possibilities to understand diseases

[5]: Yakhini et al. (2011), *Cancer computational biology*

[6]: Hasin et al. (2017), “Multi-omics approaches to disease”

such as cancer and steps forward the goal of so-called precision medicine. Precision medicine, also known as "personalized medicine", focuses on tailoring disease prevention and treatment taking into account differences in patients' genes, environments or lifestyles [7]. The final goal is to target the right treatments to the right patients at the right time, by using a data-focused approach.

To do so, researchers are focusing on the use of algorithms that are able to learn from these data, in order to solve a given task (e.g. diagnosis support for a specific cancer type). Machine learning is a field devoted to understanding and building methods that are able to leverage data to improve performance on some set of tasks [8]. The combination of machine learning algorithms with the biological data collected from cancer is revolutionizing our understanding of the disease, showing outstanding results [9]. However, the relationships and improvements that can be obtained by studying how the collected data interact and can be integrated are not yet fully exploited and should be more widely studied. By studying these relationships, better clinical-decision support systems (CDSS) can be created, helping physicians during the diagnostic process and thus, enhancing the patient prognosis.

Therefore, some basic biological concepts are outlined in this chapter in order to clarify why cancer data is increasingly requiring more and more complex computational resources, and which technologies are being used nowadays. Since the goal of this thesis is to develop multi-modal machine learning models for diagnosis and synthetic data generation, we are just going to describe the biological background necessary to understand the data used. However, we will include references to works and books where the knowledge can be further expanded.

1.2 Biological background

The cell, the DNA, and the genes

To understand the complexity of the cancer disease, it is necessary to first understand the most basic unit in biology, the cell, and its functionality. All living organisms are formed by cells, ranging from those formed by a unique cell (unicellular) to trillions of cells (like humans). The cells

[7]: Hodson (2016), "Precision medicine"

[8]: El Naqa et al. (2015), "What is machine learning?"

[9]: Kourou et al. (2015), "Machine learning applications in cancer prognosis and prediction"

contain one of the most crucial parts in biology, the genetic material, which among much other information, provides the necessary for cell replication [10]. There exist two types of cells, depending on how their nuclei are structured: prokaryotic and eukaryotic. Humans are formed by eukaryotic cells, which have a more complex structure than prokaryotic cells. They are formed by multiple compartments (which are called organelles), each one with a specific function (see Figure 1.1).

[10]: Bunz (2008), *Principles of cancer genetics*

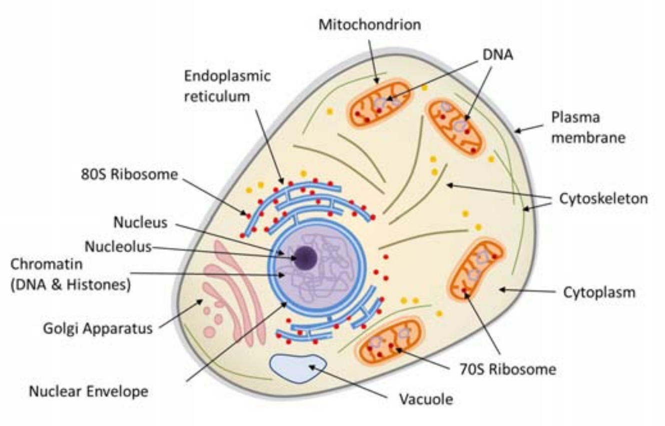


Figure 1.1: Structure and organelles of an eukaryotic cell. [10]

The plasma membrane contains the rest of the organelles, and that serves to separate and protect a cell from its surrounding environment. Inside the cytoplasm, the rest of the organelles are contained. The cytoskeleton is inside the cytoplasm, and provides the cell with its shape, anchors the organelles, and stimulates the cell movement. The ribosomes are the main site for protein synthesis and are composed of proteins and ribonucleic acids. The mitochondria, which is also known as the "powerhouse of cells", is where the energy is produced. The Golgi apparatus is where the formation of glycoproteins and glycolipids occurs. Finally, the nucleus serves as the "commander" of the cell, giving instructions for the growth, maturation, division, or death of the cell. It contains the heredity material in the deoxyribonucleic acid (DNA), which plays a key role in the cancer disease.

DNA is a polymer that holds the genetic material in almost all organisms. It contains instructions for how, when, and where to produce each kind of protein. Proteins are large biomolecules and macromolecules that comprise one or more long chains of amino acid residues, and that perform a vast array of functions within organisms. The structure of the DNA was firstly described by the Nobel prize winners James Watson and Francis Crick [11], an achievement that would have been impossible without the experimental work performed by Rosalind Franklin [12]. DNA is composed of two polynucleotide chains that coil around each other to form a double helix. The helix is composed of simpler units called nucleotides [13]. Each nucleotide is formed by one of four nitrogen-containing nucleobases: cytosine (C), guanine (G), adenine (A), or thymine (T), a sugar called deoxyribose, and a phosphate group. These bases are bound together by a hydrogen bond, the adenine with the thymine and the cytosine with the guanine. This is what joins the two strands of the DNA, forming the helix (see Figure 1.2) [14]. The order in which the bases are placed within the strands is considered to be the instruction book for building and maintaining an organism. During DNA replication, each strand serves as a template to replicate the order of the bases. Thus, in the cell division process, one strand from the original cell is copied into the new cell, and the second strand is a newly synthesized copy. The DNA strands have the opposite orientation: one strand is in the 5′ to 3′ direction with respect to the carbon atoms on the sugar (deoxyribose) and the complementary strand is in the 3′ to 5′ direction (see Figure 1.3) [15].

[11]: Watson et al. (1953), "Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid"

[12]: Sayre (2000), *Rosalind Franklin and dna*

[13]: Chaffey (2003), *Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. Molecular biology of the cell. 4th edn.*

[15]: Brenner et al. (2002), *Encyclopedia of genetics*

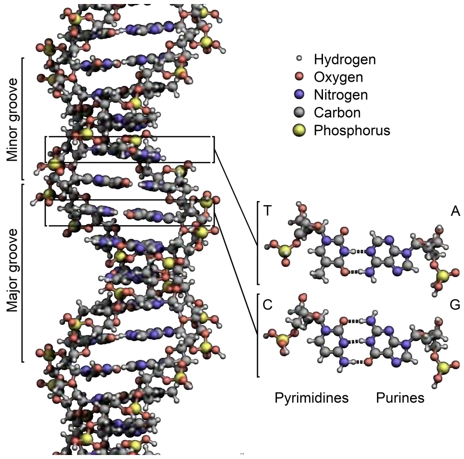


Figure 1.2: The structure of the DNA double helix. The atoms in the structure are color-coded by element and the detailed structures of two base pairs are shown in the bottom right. [14]

Within the eukaryotic cells, the DNA is organized into long structures called chromosomes. The chromosomes have a complex three-dimensional structure, that plays a key role in

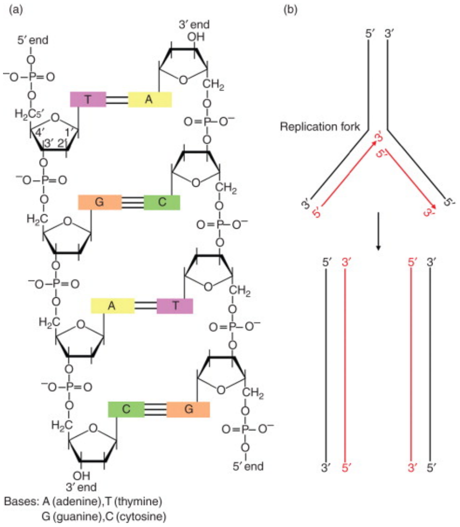


Figure 1.3: DNA strands opposite orientation from 5′ to 3′ with respect to the carbons atoms, and the opposite (form 3′ to 5′) for the other strand. [15]

transcriptomic regulation. This complex structure is called chromatin, and it contains the vast majority of the DNA of an organism its structure is maintained by the coil of DNA and proteins. When the individual chromosomes are visible, they form a classic four-arm structure, a pair of sister chromatids attached to each other at the centromere. This led the structure to have two short arms (p arms) and two longer arms (q arms) (see Figure 1.4). In humans, each cell normally contains 23 pairs of chromosomes [16].

Genes are specific units of heredity that are mapped in specific regions of the chromosome, and they mostly contain information for synthesizing specific proteins. Within each gene, we can differentiate two portions of the sequence, based on their ability to code specific amino acids of the protein (exons) or sections with non-coding capabilities (introns) [17]. The length of the genes is not fixed and can vary between hundreds to more than 2 million base pairs. The majority of the individuals in a given species share the same genes, however, there is a small portion of genes (around 1%) that provide particular heritable characteristics to the individuals (e.g. in the case of humans this could be the hair color or the body type). These characteristics are also known as traits. For learning more about the biological basics of the cell and DNA, the reader can find more information in [18, 19].

[17]: General Medical Sciences (2022), *What is genetics?*

[18]: Alberts (2017), *Molecular biology of the cell*
[19]: Cooper et al. (2007), *The cell: a molecular approach*

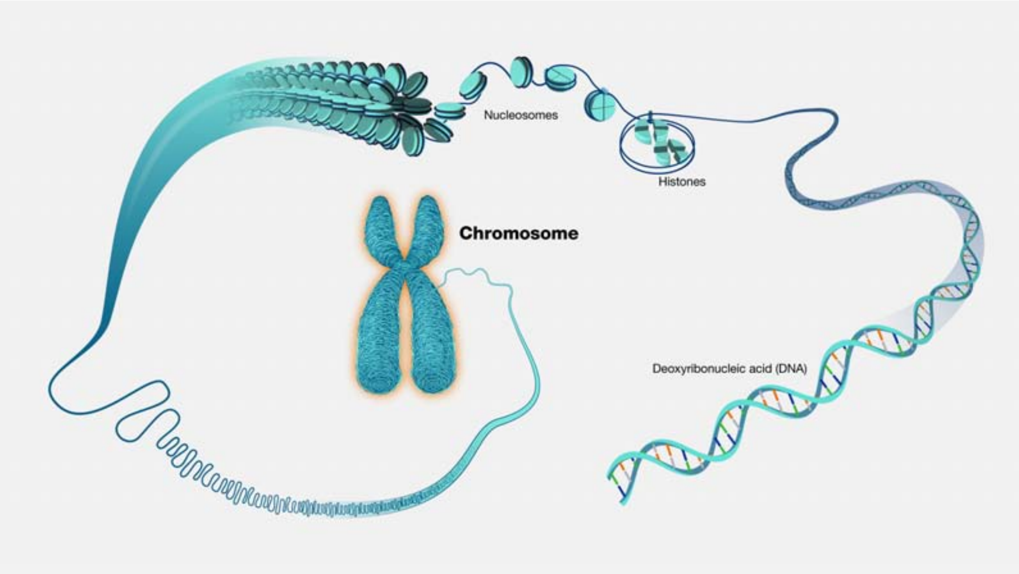


Figure 1.4: DNA is organized in chromosomes, where two short and two long arms can be differentiated [16].

The biology of cancer

Once we have understood the fundamental parts of the cell, DNA, and genes, we can focus on the cancer disease and how it develops. Cancer is a disease where cells grow uncontrollably, potentially spreading to other parts of the body. This is due to the silence of cell death. As aforementioned, human cells grow and divide to form new cells as the body needs them. When cells grow old or become damaged, these new cells replace them. However, sometimes this process does not occur, and cells that should be replaced keep growing and multiplying. These abnormal cells may grow into tumors, which are solid masses of tissue that form when abnormal cells are grouped together. There are two types of tumors: cancerous and not cancerous (also known as benign). While benign tumors may imply some danger based on their location and size, usually they can be safely removed and they do not grow again nor invade other parts of the body. On the other hand, cancerous tumors can spread to nearby tissues and can travel to others far apart in a process called metastasis, and once removed they can grow back [20, 21].

Cancerous cells have several differences from normal cells. Firstly, they grow in the absence of growth signals. They also ignore the death signaling, which tells cells to stop

[20]: Pecorino (2021), *Molecular biology of cancer: mechanisms, targets, and therapeutics*
[21]: Institute (2022), *What is cancer?*

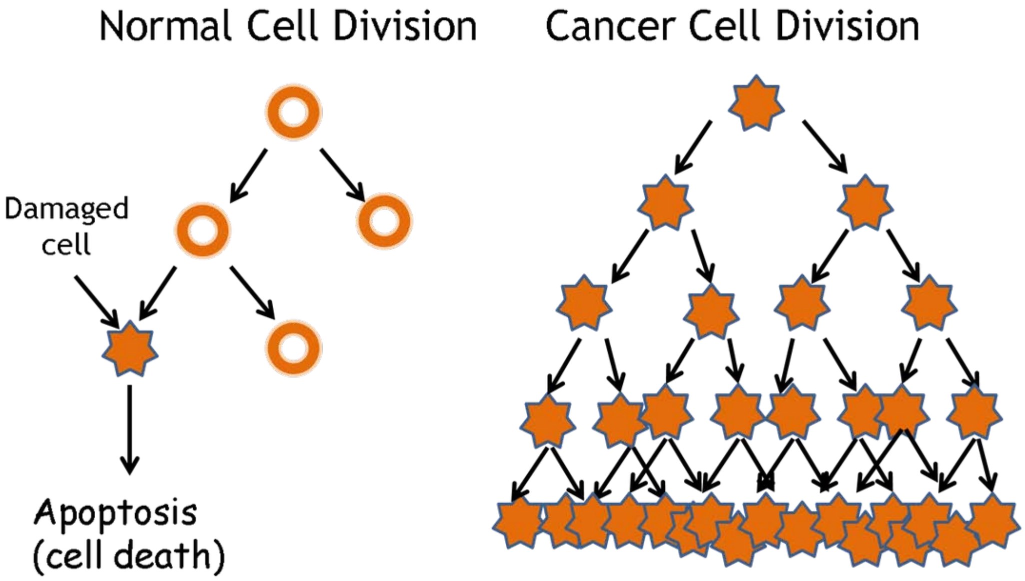


Figure 1.5: In a normal cell division, if a cell is damaged the process of apoptosis (cell death) triggers. With cancerous cells, this does not occur, and damaged cells keep dividing. [22]

dividing (a process called apoptosis). Cancerous cells are able to hide from the immune system, which normally eliminates damaged or abnormal cells. Not only that, but sometimes they also trick the immune system to help them grow, convincing the immune system to protect these cancerous cells instead of eliminating them. Finally, they also accumulate changes in their chromosome, such as duplications and deletions of some parts [20, 21].

Cancer is a genetic disease since it is caused by changes that affect those genes controlling the division and growth of the cell. These changes can be initiated by different factors, such as errors during cell division, environmental factors that affect our genome (such as smoking or pollution), or inherited traits from our parents. While the immune system usually eliminates the damaged DNA before it turns cancerous, this ability deteriorates as we age. Therefore, we have a higher chance of developing cancer in a later stage of our life.

Cancer is usually developed by genetic changes in three main types of genes: proto-oncogenes, tumor suppressor genes, and DNA repair genes. Proto-oncogenes are normal genes that are involved in normal cell growth. However, if these genes have been modified in some way or are more active than normal, they can turn into cancer-causing genes (also known as oncogenes), causing the abnormal growth of cells and their survival when they should be eliminated. Similarly, tumor suppressor genes are also involved in cell division and growth. Those cells that present alterations in tumor suppressor genes divide uncontrollably. Lastly, DNA repair genes are in charge of repairing damages in the DNA that can occur during cell division or by other external factors. Those cells with mutations in these genes tend to present deletions or insertions in parts of their chromosomes, affecting other genes. These mutations, or the accumulation of these, can turn the cell cancerous [20, 21].

Contrary to popular belief, cancer is not a single disease, but a group of them that share the same root. There are more than 100 cancer types, which are usually named by the organ or tissue where they formed. For instance, breast cancers originated in the breast or lung cancers on the lung. On the other hand, cancer can also be named based on the type of cell that formed them, such as the epithelial cell or, for instance in the case of lung cancer, the squamous cell [21]. To get a deeper description of the cancer disease, the reader can go to the following references [23, 24].

[23]: Ruddon (2007), *Cancer biology*
[24]: Mader (2007), "The biology of cancer"

Multi-Omics

In 2001 we obtained the first sequence of the human genome [25], covering 93% of it, marking a milestone in human biology research. Only recently a complete sequence of the human genome has been attained [26]. This huge achievement has been possible thanks to technological advances, making it possible to measure different aspects of a tissue or cell biology with high quality. With the modernization of instrumentation, different aspects of the biology of tissues or cells have been more accurately measured. Thus, different scientific fields have been created depending on which aspects are being measured, called "omics" [27]. The use of a multi-omics approach allows us to comprehensively understand the biological subject that is being studied. There are multiple omics, and with the recent advances in information technologies more are being developed. Examples include genomics, transcriptomics, epigenomics, proteomics, and metabolomics, which study genes, RNA, methylated DNA, proteins, and metabolites respectively, just to mention a few

[25]: Venter et al. (2001), "The sequence of the human genome"

[26]: Nurk et al. (2022), "The complete sequence of a human genome"

[27]: Micheel et al. (2012), "Omics-based clinical discovery: Science, technology, and applications"

(see Figure 1.8). Here we are going to describe those that have been involved in the development of this thesis.

Genomics is the field that studies the human genome. To study it, researchers sequence the base pairs that form the strands of the DNA. One of the most used technologies is dye sequencing, firstly developed by Shankar Balasubramanian and David Klenerman, and now acquired by the company Illumina. This method is based on reversible dye-terminators that enable the identification of single nucleotides as they are washed over DNA strands (see Figure 1.6). Firstly, the DNA is purified. Then, the DNA is fragmented and adapters are integrated, which would help with the next steps. The DNA is loaded onto a flow cell, where the amplification and sequencing are going to be carried out. The flow cell contains nanowells that space out fragments and help with overcrowding. Each one of the nanowells contains oligonucleotides that can attach to the previously added adapters. Once these adapters have been attached, the cluster generation phase begins. A thousand copies of each fragment of DNA are obtained using the polymerase chain reaction (PCR) method. Then, primers and modified nucleotides are washed onto the chip. The particularity of these nucleotides is that they have a reversible fluorescent blocker, only allowing the DNA polymerase to add one nucleotide onto the DNA fragment each time [28]. After each round of synthesis, a picture of the chip is taken by a camera. Then, using a computer determines which base has been added based on the wavelength of the fluorescent tag, and saves it for every spot on the chip. After that, the remaining molecules are washed away and fluorescent terminal blocking groups are removed. This process is repeated until the full DNA molecule is sequenced [29]. This process can be fully paralleled, allowing the sequence of thousands of places throughout the genome using massive parallel sequencing, also called next-generation sequencing (NGS) [30]. Having the whole-genome sequence allows us to study variations (known as mutations) with respect to a control genome. This can be substitutions, deletions, or insertions of parts of segments of the genome, and can lead to the development of diseases. New sequencing technologies are being introduced every year, with their advantages and disadvantages. If the reader wants to learn more about sequencing technologies, they can go to the following works [31–34].

Transcriptomics is the study of the transcriptome which is un-

[28]: Clark et al. (2013), *Molecular biology*

[29]: Meyer et al. (2010), "Illumina sequencing library preparation for highly multiplexed target capture and sequencing"

[30]: Behjati et al. (2013), "What is next generation sequencing?"

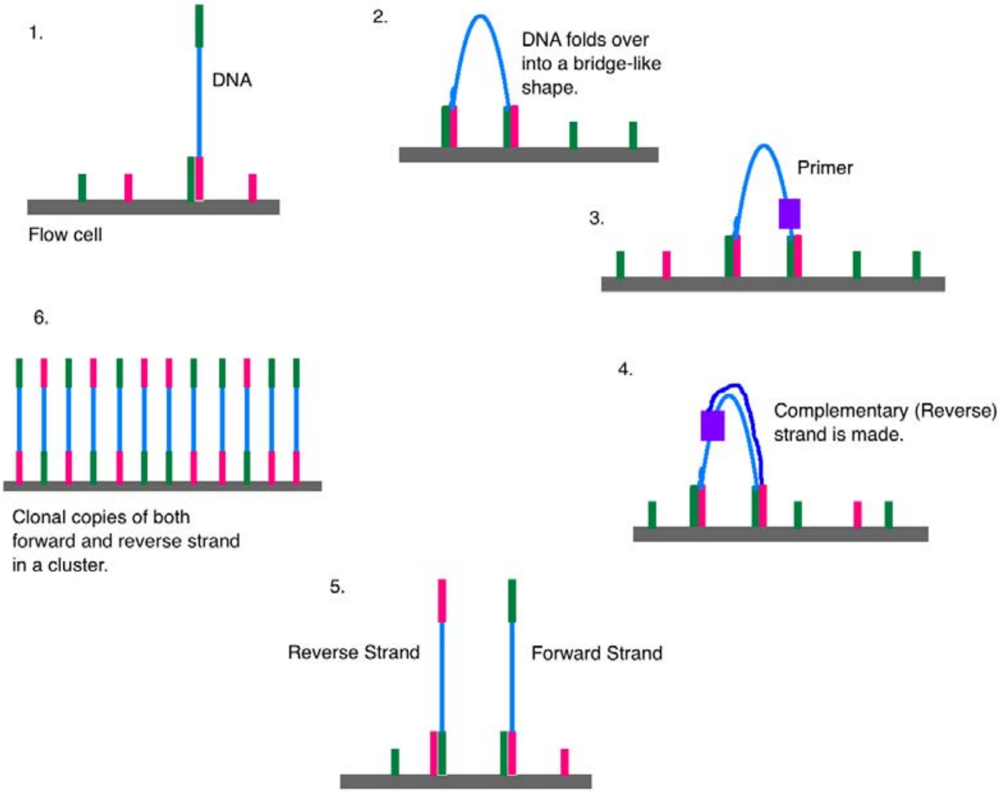

1.
DNA
Flow cell
2.
DNA folds over into a bridge-like shape.
3.
Primer
4.
Complementary (Reverse) strand is made.
5.
Reverse Strand
Forward Strand
6.
Clonal copies of both forward and reverse strand in a cluster.

Figure 1.6: The DNA attaches to the flow cell via complementary sequences. The strand bends over and attaches to a second oligonucleotide forming a bridge. A primer synthesizes the reverse strand. The two strands release and straighten. Each forms a new bridge (bridge amplification). The result is a cluster of DNA forward and reverses strand clones. [35]

derstood as the complete variety of ribonucleic acids (RNAs) that are expressed in a cell, tissue, or organism [36]. RNAs are important macromolecules that are produced, based on the DNA, by the cellular process of transcription (the process in which DNA is translated to RNA). All the different types of transcripts are covered in transcriptomics, such as messenger RNAs (mRNAs), microRNAs (miRNAs), ribosomal RNAs (rRNAs), or non-coding RNAs (ncRNAs). However, in humans, the transcriptome of the protein-coding genes only represents between 1.5 and 2 percent of the genome. There are two major technologies that allow us to measure the transcriptome: microarrays and RNA-Seq. Microarrays are oligonucleotide-based probes, similar to the dry sequencing for genome data, that hybridize into specific RNA transcripts. RNA-Seq is more recent and advanced, allowing to directly sequence RNA transcripts without the need for probes (see

[36]: Bradshaw et al. (2015), *Encyclopedia of cell biology*

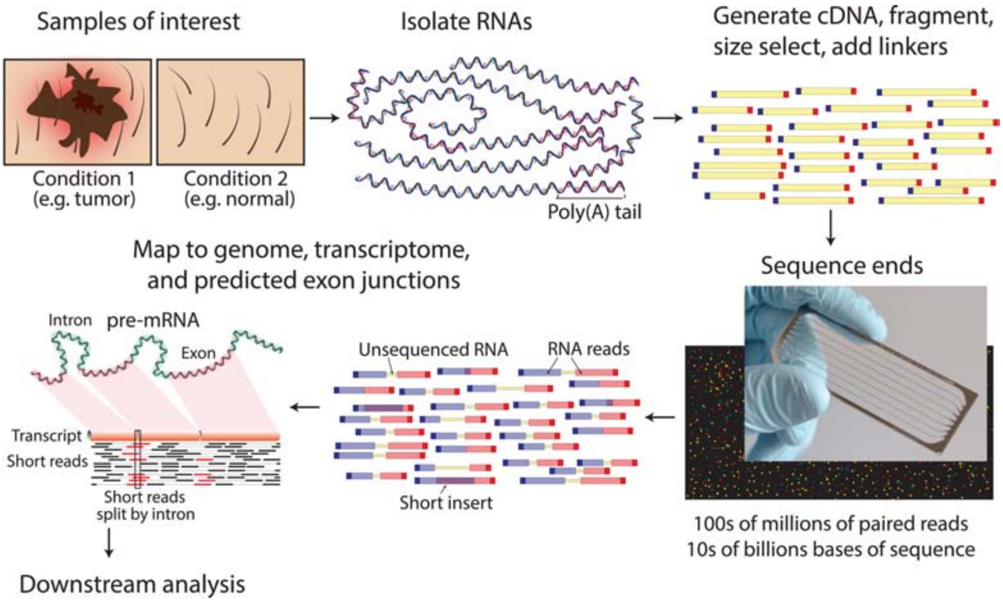

Samples of interest
Isolate RNAs
Generate cDNA, fragment, size select, add linkers
Condition 1 (e.g. tumor)
Condition 2 (e.g. normal)
Poly(A) tail
Map to genome, transcriptome, and predicted exon junctions
Sequence ends
Intron
pre-mRNA
Exon
Unsequenced RNA
RNA reads
Transcript
Short reads
Short reads split by intron
Short insert
100s of millions of paired reads
10s of billions bases of sequence
Downstream analysis

Figure 1.7: Several steps are involved in a typical RNA-Seq workflow. First, the RNA samples of interest are isolated. Then, sequencing libraries are generated. A high-throughput sequencer is used to produce hundreds of millions of short paired-end reads. These reads are aligned against a reference genome of transcriptome and finally, downstream analysis is carried out for the expression estimation, differential expression, or other applications. [39]

Figure 1.7). Once we have measured the RNA, we can detect changes in gene expression, helping to identify biomarkers for specific diseases. More information about how transcriptomics technologies work and how they can help in fighting diseases, can be found in these two books [37, 38].

Lastly, epigenomics consists of measuring reversible chemical modification of the DNA, that produced changes in the expression of the genes without modifying the original base sequence. Epigenomics modifications can occur for environmental factors that affect or in the development of disease states. Environmental factors can be external (such as pollution) or patient behaviors (such as smoking tobacco) [40, 41]. Biochemically, epigenetic changes that are measured at high throughput belong to two categories: methylation of DNA cytosine residues (at Cytosine-Guanine sites (CpGs)) and multiple kinds of modifications of specific histone proteins in the chromosomes (histone marks). A CpG site is just a point of the genome where a cytosine is followed by a guanine in the 5′->3′ direction of the strand. It controls the expression of those genes that are close to the CpG site. To test if a CpG site is methylated, the genome is treated with bisulfite. If it is methylated, it will stay as a CpG site, otherwise, the cytosine

[40]: He et al. (2018), "Role of genetic and environmental factors in DNA methylation of lipid metabolism"

[41]: Gao et al. (2016), "Tobacco smoking and methylation of genes related to lung cancer development"

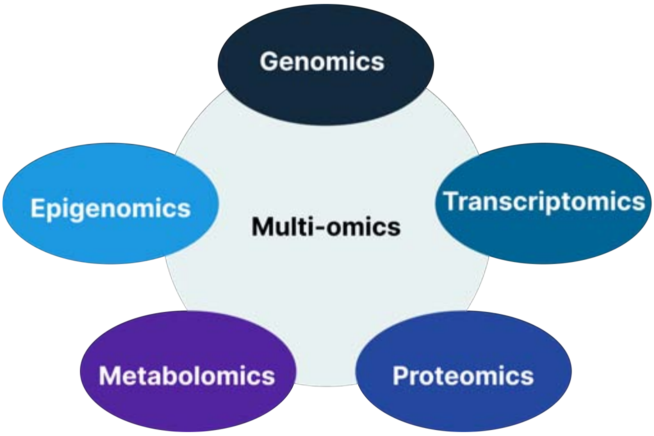


Figure 1.8: Different fields inside the multi-omic approach. Genomics studies the genome, transcriptomics studies the transcriptome, epigenomics studies the measure modification in the DNA, proteomics studies the proteins, and metabolomics the metabolites.

will change to a thymine [42]. To deepen the understanding of epigenomics, the reader can refer to the following book [43] and, specifically for cancer, the following review [44].

[42]: Yang et al. (2004), "A simple method for estimating global DNA methylation using bisulfite PCR of repetitive DNA elements"

Digital Pathology

Digital pathology is a sub-field of pathology that is focused on the organization of digitized tissue slides, thanks to the use of virtual microscopy. Then, these slides can be viewed, managed, shared and analyzed on a computer monitor [45]. The most used digital pathology tool nowadays is whole-slide imaging (WSIs). A WSI scanner is a robotic microscope that is able to digitize an entire glass slide, stitching high-resolution individually captured images to generate an even higher-resolution image of the whole slide. Once the file has been captured, it can be viewed, magnified, or investigated through the computer just as you could do with a traditional microscope [46]. The first WSI scanners were introduced in the 90s, and they were less advanced than their counterparts at the time [47]. However, after the introduction to the market of an accurate, fast, and cost-efficient scanner by Interscope Technologies [48], they situated WSI as the state-of-the-art for digitized tissue slides [49]. There are two main approaches for producing digital images. The majority of the available machines use a tiling system, where different tiles are obtained from the original images while others employ line-scanning systems, capturing the tissue in a linear way [49] [50]. Nowa-

[45]: Pantanowitz et al. (2018), "Twenty years of digital pathology: an overview of the road travelled, what is on the horizon, and the emergence of vendor-neutral archives"

[46]: Hanna et al. (2020), "Whole slide imaging: technology and applications"

[47]: Pantanowitz et al. (2011), "Review of the current state of whole slide imaging in pathology"

[48]: Ogilvie (2005), *Virtual microscopy and virtual slides in teaching, diagnosis, and research*

[49]: Farahani et al. (2015), "Whole slide imaging in pathology: advantages, limitations, and emerging perspectives"

[50]: Evans et al. (2017), "Implementation of whole slide imaging for clinical purposes: issues to consider from the perspective of early adopters"

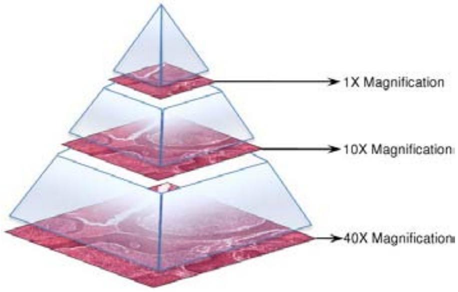


Figure 1.9: Example of the pyramidal structure of WSI. The different magnifications are stored, from the highest magnification (40x) to the lowest (1x). [55]

days, scanners have the ability to produce digitized slides in the span of minutes (or less) [49]. Different scanners have different features or capabilities. For instance, the scanning capacity can vary between 100-200 slides in a single batch, the objective availability (usually 20x or 40x magnifications), and the image resolution (0.25-0.5 μm per pixel). WSIs have been used for a wide variety of tasks in clinical environments, including telepathology for primary diagnosis, consulting other physicians at other hospitals or remotely interpreting frozen sections [49].

Once the WSIs have been obtained, pathologists examine them to find features that can help them to perform a diagnosis. In order to facilitate this task, and especially for cancer, tissues are stained with hematoxylin and eosin (H&E) stain. In this case, the hematoxylin stains cell nuclei in a purplish blue, and the eosin stains the extracellular matrix and the cytoplasm in pink, with the structures taking different hues and shades (see Figure 1.10) [51]. This stain allows physicians to detect abnormalities in the tissue, that could indicate the development of cancer. We refer to the following works [52–54] for a more in-depth review of all the possibilities that WSIs offer in pathology.

[51]: Chan (2014), "The wonderful colors of the hematoxylin–eosin stain in diagnostic surgical pathology"

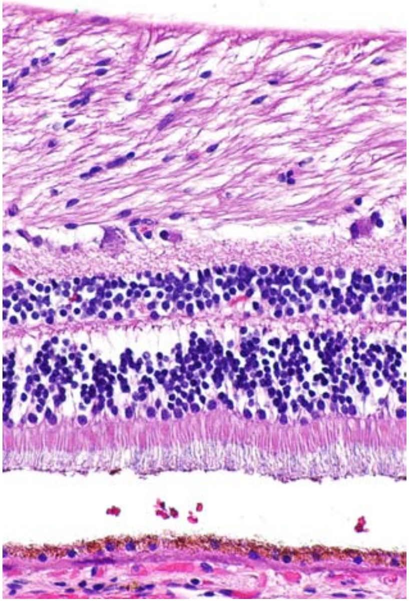


Figure 1.10: Retina tissue stained with hematoxylin and eosin, where cell nuclei are stained blue-purple and the extracellular material in pink. [56]

1.3 Publicly available multi-scale and multi-omics databases

During this thesis we are going to refer to the concept of multi-scale/multi-omics as related ones, to also include histopathology as one of the main methodologies for cancer identification. Given the multi-scale nature of cancer data, and biology in general, great efforts have been made to build multi-omic and multi-modal databases. One of the most important databases for cancer research is The Cancer Genome Atlas (TCGA) project [57]. TCGA contains information from 33 different cancer types, and they have achieved the goal of providing easy access to the data, through the GDC platform. Not only that, but they have harmonized the data, making it easy to use and analyze. Multiple modalities are available for each patient (e.g: RNA-Seq, WSIs, miRNA-Seq, DNA Methylation, Copy Number Variation, Single-Nucleotide-Polymorphisms, etc.), enabling the study of the disease in a multi-scale way. Having this variety greatly helps the task of creating multi-modal CDSS, and also researching the interactions between different omics. All the information about the project can be

[57]: Weinstein et al. (2013), "The cancer genome atlas pan-cancer analysis project"

read in their webpage [*].

More databases are being created with the same approach in mind, gathering all the possible information from the same patient. The Genotype-Tissue Expression (GTEx) project is an ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation [58]. Samples were collected from 54 non-diseased tissue sites across nearly 1000 individuals, primarily for molecular assays including WGS, WES, and RNA-Seq. This allows studying human biology in a non-disease status. More information and the data can be accessed in the GTEx project webpage [†].

[58]: Lonsdale et al. (2013), "The genotype-tissue expression (GTEx) project"

Other databases now include more multi-modal information, even though it was not the case at the beginning. The Cancer Imaging Archive (TCIA) is a service which de-identifies and hosts a large archive of medical images of cancer accessible for public download [59]. They contain imaging data in multiple modalities (MRI, CT, digital histopathology, etc.), but now, they also include other information related to the patient when available, such as patient outcomes, treatment details, genomics, and expert analyses. More information and access to the data can be obtained in their webpage [‡].

[59]: Clark et al. (2013), "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository"

1.4 Non-small-cell lung cancer

Along the first half of the research results presented in this thesis, we focused on one of the deadliest cancer types, lung cancer. We selected this cancer type based on its importance and the number of publicly available samples. It is one of the most frequent cancer types, with a total of 2.2 million new cancer cases and 1.8 million deaths worldwide in 2020 [4], representing 18.0% of total cancer related deceases. Lung cancer is characterized by uncontrolled cell growth in tissues of the lung organ. Most of the cancers that start in the lungs are carcinomas [60], and two main types can be differentiated within them: small-cell lung carcinoma (SCLC) representing around 15 - 20% of lung cancer cases, and non-small-cell lung carcinoma (NSCLC) representing around 80 - 85% of lung cancer cases [61] (see Figure 1.11). Within NSCLC,

[4]: Sung et al. (2021), "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries"

[60]: Organization (2014), *World Cancer Report 2014*

[61]: UK (), *Types of lung cancer*

[*] `https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga`

[†] `https://gtexportal.org/home/`

[‡] `https://www.cancerimagingarchive.net/`

two main different subtypes can be differentiated, which are adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC). Other minority types are also considered under the umbrella of NSCLC, such as large-cell-carcinoma and more poorly differentiated variants [62]. LUAD usually comes from peripheral lung tissue and it is usually associated with lifelong nonsmokers [63]. On the other hand, LUSC is closely correlated with a history of tobacco smoking and tended to be more often centrally located [64, 65].

[62]: Goldstraw et al. (2011), "Non-small-cell lung cancer"

[63]: Subramanian et al. (2007), "Lung cancer in never smokers: a review"

[64]: Kenfield et al. (2008), "Comparison of aspects of smoking among the four histological types of lung cancer"

[65]: Travis et al. (1995), "Lung cancer"

Accurately detecting the NSCLC subtype is a crucial task, given that treatments differ between them and the effect of the treatment has a direct impact on the patient's prognosis [62]. The variables that can be associated with prognosis are usually grouped into different categories: tumor-related, patient-related, and environmental factors [66]. Tumor-related factors are the cell type, the primary site of the tumor, or the extension of the disease. Patient-related are factors sex, comorbidity, or performance status. Lastly, environmental factors include the nutrition of the patient or the choice and quality of the treatment.

[62]: Goldstraw et al. (2011), "Non-small-cell lung cancer"

[66]: Gospodarowicz et al. (2003), "Prognostic factors in cancer"

The diagnostic process usually begins when the patient presents suspicious symptoms and signs. These symptoms can vary, but among them, we can usually find a persistent cough that does not go away after 2 or 3 weeks (even reaching the point where the patient is coughing up blood) persistent tiredness or lack of energy, or persistent breathlessness. Clinicians first used image methods, such as CT scans, to locate the lesion and then they select the appropriate technique to obtain a biopsy sample to confirm a pathological diagnosis [67]. This pathological diagnosis can be performed by means of the aforementioned biological sources. Mainly, a first diagnosis can be performed using the tissue slide [68]. However, distinct mutational subtypes can be found in pathways and receptors, as we further explore in Chapter 4. The reader can find more information about lung cancer, its prognosis, and treatments in the following references [69–72].

[68]: Travis (2002), "Pathology of lung cancer"

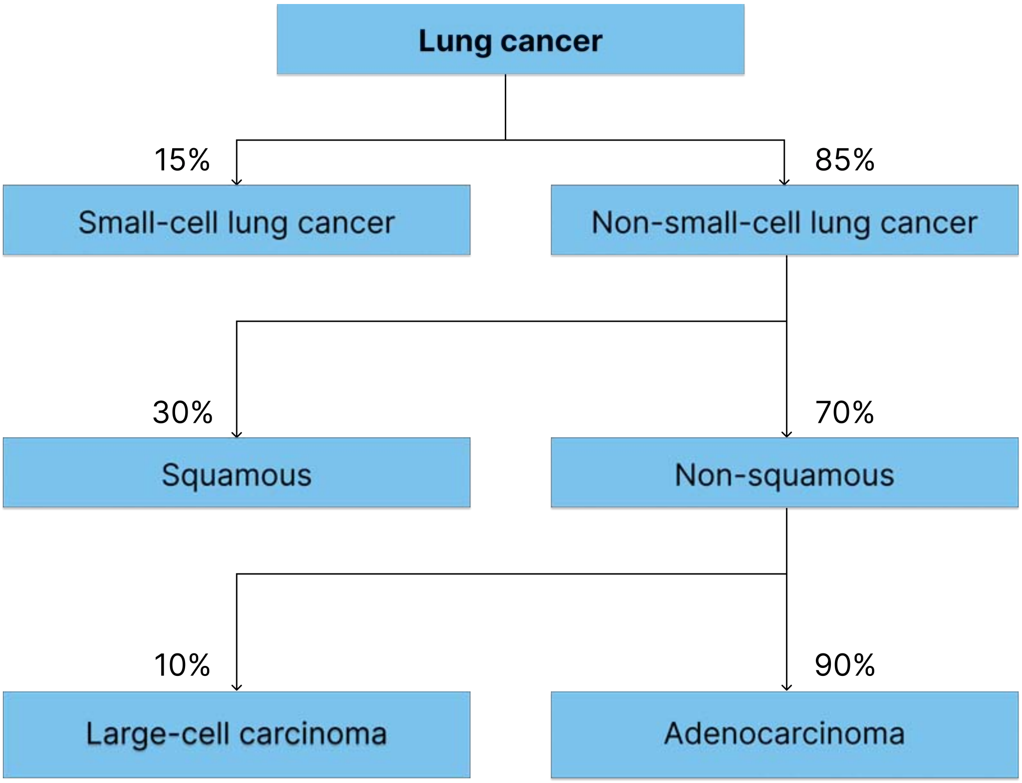


Figure 1.11: Different lung cancer subtypes and in which percentage they are found. We have focused on the two more prevalent ones inside NSCLC, squamous and adenocarcinoma [69].

1.5 Conclusions

In this chapter, we have introduced the main concepts of biology, specifically, cancer biology. The advances that are being made in the field of multi-omics thanks to the development of cheaper and faster technologies have been explained, showing the potential for the creation of multi-modal machine learning models. Then, lung cancer biology has been briefly explained, given that it is the cancer type this thesis has focused more on. In the next chapter, we will introduce the basics of machine learning, along with the techniques that have been explored during the development of this thesis.

Fundamentals of machine learning 2

2.1 Introduction

Machine learning (ML) is a field of artificial intelligence in which, by using mathematical models, a computer is able to learn from the available data to solve a given task. Machine learning uses algorithms to identify patterns within data, and those patterns are then used to create a data model that can make predictions [73]. Within ML, we can differentiate three areas, depending on the nature of the feedback available to the learning system: supervised, unsupervised or reinforcement learning [73]. In this thesis, we will focus on supervised and unsupervised learning, but we will briefly explain reinforcement learning so the reader has the full picture. If they want a more detailed explanation, they can refer to [74].

In supervised learning, we try to model a function, $f(x)$, based on pairs of data samples and their desired outputs [75]. Usually, these pairs are called training data. Generally speaking, by an iterative optimization of the objective function using the training data, the model learns a function, $\hat{f}(x)$, that can predict the output of a given input. These outputs can be anything, from categories to real values. Depending on the type of variable that we want to predict, two groups of problems can be discerned: classification and regression problems. In a classification problem, we want to predict a categorical value, such as cancer type. On the other hand, in a regression problem, the output is a real value (e.g. life expectancy in a prognostic prediction problem). Some specific techniques are applied to each kind of problem, but most well-known modeling techniques have alternatives that can be used for either category.

Unsupervised learning, unlike supervised learning, only has non-categorized data samples and aims to learn the structure of the data. Typical examples of this include the application of clustering techniques to find hidden groups based on characteristics or finding the probability density function (PDF) in order to be able to sample (or generate) new data points [75]. Examples of unsupervised learning problems in biomedicine are finding groups of patients that respond similarly to the same treatment, or generating synthetic gene expression data by estimating the probability density function of the underlying data.

Finally, reinforcement learning (RL) is an area of machine learning more closely related to robotics, where software agents take actions in an environment trying to maximize some notion of cumulative reward based on the actions [76]. By maximizing that reward, the agent is able to learn which actions to take with respect to the environment. RL has been notoriously used to create agents that are able to play video games, even surpassing humans [77–79].

[76]: Otterlo et al. (2012), "Reinforcement learning and markov decision processes"

[77]: Silver et al. (2016), "Mastering the game of Go with deep neural networks and tree search"
[78]: Silver et al. (2017), "Mastering the game of go without human knowledge"
[79]: Vinyals et al. (2019), "Grandmaster level in StarCraft II using multi-agent reinforcement learning"

Below, we are going to introduce the basics of some of the key machine learning models that have been used in this thesis. For a more detailed description of the presented techniques, including all the mathematical background, the reader can consult the following resources [73, 80, 81]

[73]: Bishop et al. (2006), *Pattern recognition and machine learning*
[80]: Robert (2014), *Machine learning, a probabilistic perspective*
[81]: Murphy (2012), *Machine learning: a probabilistic perspective*

2.2 Artificial Neural Networks and Convolutional Neural Networks

Artificial Neural Networks (ANNs) are learning algorithms based on the functioning of biological neural networks, and they can be used under any ML paradigm. The most general type of ANN is the multi-layer perceptron, also known as Feed Forward Neural Network (FFNN). In this type of network, the information moves in only one direction, forward, through the different layers [82]. The basic building blocks of ANNs (and FFNNs specifically) are the so-called neurons. The neuron is formed by a weight vector W, a unique value named bias b, and the activation function. The neuron calculates the inner product of its inputs and the weight vector plus the bias and, given this calculus, the activation function determines whether the neuron will activate or not. Neurons are grouped in layers, which in the case of FFNNs can be grouped in three main categories: input, output, and hidden layers. The input layer is the one receiving the input data. In the case of an MLP, we will have as many neurons in this layer as the input data dimension. Then, the input layer can be followed by one (or multiple) hidden layers. These layers process the information from a previous layer and calculate the output (that, subsequently, serves as input for the next

[82]: Schmidhuber (2015), "Deep learning in neural networks: An overview"

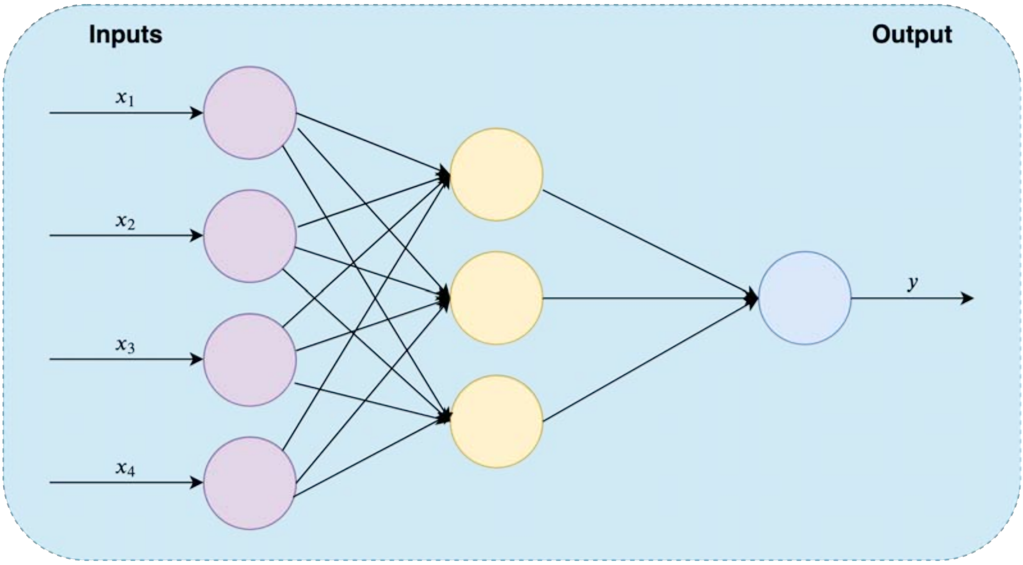


Figure 2.1: Example architecture of FFNN with four inputs, one hidden layer with three neurons, and one output [85].

layer). Finally, we have the output later, which determines the output of the network. We will have as many neurons as the dimensionality of the desired output (e.g. one neuron if we are doing prognosis prediction or the number of classes in a pancancer classification problem).

For a given problem, neuron weights need to be learned in order to properly approximate the function that would classify or predict an input. To accomplish this task, the backpropagation algorithm [83] was proposed. It uses a dataset of samples of a problem with known output, and in an iterative way updates the weights as follows:

[83]: Rumelhart et al. (1988), "Learning representations by back-propagating errors"

$$\boldsymbol{w}^{k+1} = \boldsymbol{w}^{k} + \triangle\boldsymbol{w} \tag{2.1}$$

$$\triangle\boldsymbol{w}_i = -\eta \cdot \frac{\partial E}{\partial \boldsymbol{w}} \tag{2.2}$$

where η is the learning rate and $\frac{\partial E}{\partial w}$ is the gradient of the error with respect to the weights. The gradient gives how a function varies with respect to the variable being derived. The negative sign is used since the error needs to be minimized. FFNNs were proven to be a universal approximator [84]. A basic scheme of an FFNN can be observed in Figure 2.1.

[84]: Hornik et al. (1989), "Multilayer feedforward networks are universal approximators"

Based on the popularity of the multi-layer perceptron, other neural network architectures, also with a severe impact in ML, have been proposed. One example is CNNs, with a strong impact on Computer Vision thanks to an architecture based on the convolution operation, which is applied as a matrix multiplication between a filter and the data. CNNs have been around for a long time. Although they were first proposed in 1980 [86], it was not until diverse modifications were applied to the learning algorithms, the quantity of data available was dramatically increased, and the necessary computing platforms were developed, that CNNs were revisited. One of the most important works from this new era of CNNs was proposed for solving a digit classification problem [87]. Since then, many studies have been published showing CNNs outperforming other established techniques (in some cases even outperforming human capabilities) in multiple computer vision problems [88], including pattern recognition, image segmentation, and image generation.

What makes CNNs so special is their refined ability to automatically extract features from data. Previously, features needed to be extracted by hand from images for later processing [89]. Due to its complexity, this was considered one of the toughest tasks in computer vision. To automatically extract features, the convolution operator is used (which applies a defined matrix of numbers to the input). At its most basic, it can easily detect edges, lines, textures, and other simple patterns in an image. By using different layers together with an adequate learning algorithm, more complex filters can be learned. These more complex filters would not only be able to detect specific complex shapes in images or in the signals presented, which are more relevant to the problem tackled, but they would also improve the model performance when compared with other more traditional methods. In fact, in CNNs, by applying the convolution operation and through the backpropagation of the error for the weights update, those complex filters are learned in a straightforward manner [90]. By grouping the convolutional operation in convolutional layers, different specific features can be learned within the same layer. The use of multiple layers would lead to a hierarchical structure where the first layers will learn basic features (e.g. lines or borders) and will pass that information to the remaining layers in order to detect more complex features (e.g. cell shapes and tumor heterogeneity).

CNNs are formed by different kinds of layers, the main ones are the convolutional, the pooling, and the fully connected/-

[86]: Fukushima (1980), “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position”

[87]: LeCun et al. (1998), “Gradient-based learning applied to document recognition”

[88]: LeCun et al. (2015), “Deep learning”

[89]: Harris et al. (1988), “A combined corner and edge detector.”

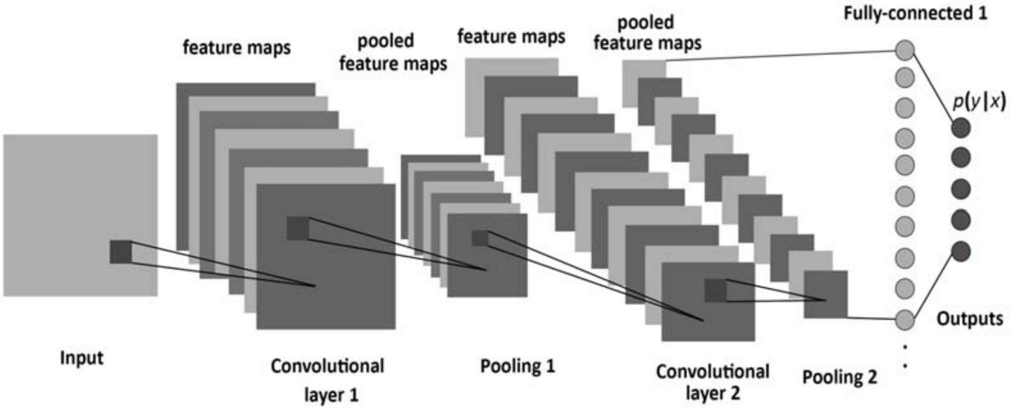


Figure 2.2: Example of a general CNN architecture [91].

dense layers. The convolutional layers, as their name implies, apply the convolution operation to the data spatial domain. Their main goal is to extract the information available in the data. Internally they are weight matrices, which are optimized and learned during the training process. Pooling layers are used as dimensionality reduction techniques, combining the outputs of multiple neurons into a single one in the next layer. Usually, they are placed after the activation function, and they increasingly reduce the amount of information extracted by the convolutional layers [90]. Finally, a set of fully connecteddense layers is used where the learned features are the input and the output is the final task goal.

A general architecture for a CNN can be observed in Figure 2.2. The feature extraction operation is performed by a set of convolutional filters (in the form of convolutional layers). Later, those filters are used for predictive tasks, using fully connected layers.

An important aspect to be considered when using DL techniques is that they normally require large databases of high-quality images to learn very specific patterns. This requires much computational time and powerful HPC systems. In order to avoid these possible drawbacks, transfer learning (TL) allows the use of a pre-trained network (a DNN that has been trained on another dataset), which is able to identify complex patterns in image data, for a certain application. That network could be partially retrained (fine-tuning) with the application database (which is usually much smaller) in order to classify a set of moderately different patterns. This greatly expands the usability of DL models for specific tasks, by learning global image patterns in a sufficiently big database, but refining them with a smaller, more specific one. Several networks are commonly used by researchers to approach computer vision problems using TL, such as GoogLeNet Inception v3 [92], ResNet [93] in its various forms (Resnet-18, Resnet-34, Resnet-50, Resnet-101 and Resnet-152) or VGG net [94]. Details and examples of this general technique can be reviewed in several published papers [95–97].

[92]: Szegedy et al. (2016), "Rethinking the inception architecture for computer vision"

[93]: He et al. (2016), "Deep residual learning for image recognition"

[94]: Simonyan et al. (2014), "Very deep convolutional networks for large-scale image recognition"

[95]: Lu et al. (2015), "Transfer learning using computational intelligence: A survey"

[96]: Weiss et al. (2016), "A survey of transfer learning"

[97]: Yang et al. (2020), *Transfer learning*

2.3 Generative Adversarial Networks

Generative adversarial networks (GANs) are an ANN framework based on a game theory scenario where two players (the generator network and the discriminator network) play against each other, firstly introduced by Goodfellow et al. [98]. The generator network produces samples based on what it learns from the training data, while the discriminator network tries to distinguish between samples drawn directly from training data and those produced by the generator. The discriminator emits a probability for that sample being drawn from training data or produced by the generator. Therefore, the discriminator's goal is to correctly classify samples as real or fake. At the same time, the generator tries to fool the classifier into believing its samples are real, learning from the data presented. At convergence, the generator's samples are indistinguishable from real data, and the discriminator outputs everything as real data. A diagram of a GAN can be observed in Figure 2.3. To optimize the parameters of both networks, different loss functions have been proposed in the literature. The first one, proposed in the original paper, is the Min-Max loss function. Both networks (generator and discriminator) are using the same loss, but the discriminator is trying to minimize the loss and the generator tries to maximize it. The loss is depicted below:

[98]: Goodfellow et al. (2014), "Generative adversarial nets"

$$\mathbb{E}_x[\log(D(x))] + \mathbb{E}_z[1 - D(G(z))] \tag{2.3}$$

where $D(x)$ is the probability outputted by the discriminator of that a real sample is real, $G(z)$ is the generator output when input the noise z and $D(G(z))$ is the discriminator probability for a fake sample being real. This formula is derived from the cross-entropy loss function.

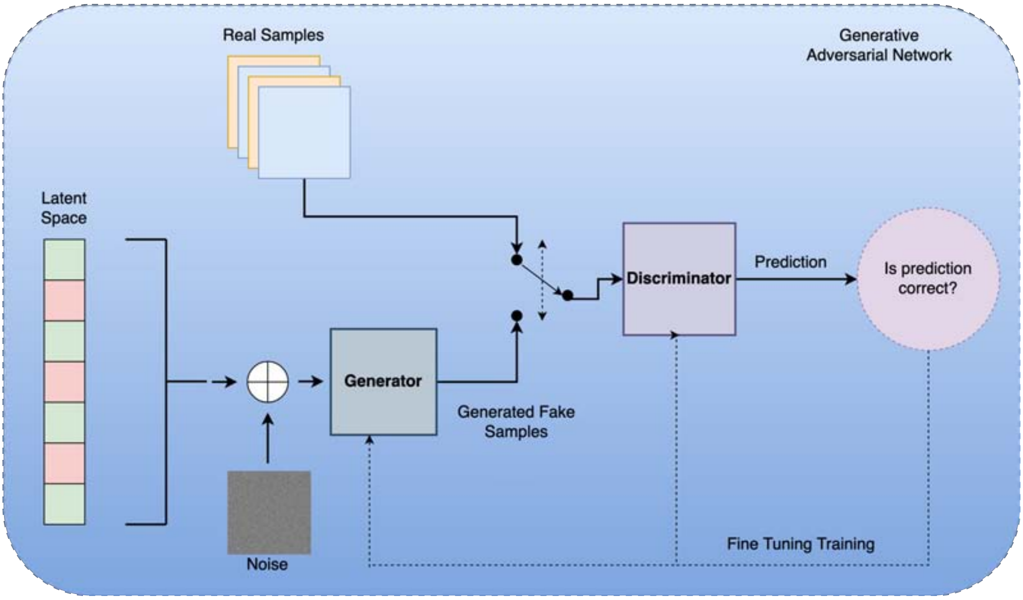


Figure 2.3: Diagram of a generative adversarial network [85].

However, this loss can produce a model collapse (the generator only produces one kind of sample) and it leads to unstable training. Therefore, new loss functions were presented in the literature, such as the Wasserstein loss by Arjovsky et al. [99], also adding to it the gradient penalty proposed by Gulrajani et al. [100]. In this case, the discriminator (or critic, as called in the paper) does not classify between real and synthetic samples, but for each sample, it outputs a number. The discriminator training just tries to make the output bigger for real samples and smaller for synthetic samples. This simplifies the loss function of both networks, where the discriminator tries to maximize the difference between its output on real instances and its output on synthetic instances as follows:

$$L_D = \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}[D(\tilde{x})] - \mathbb{E}_{x\sim\mathbb{P}_r}[D(x)] + \lambda\mathbb{E}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}[(\|\nabla_{\hat{x}}D(\hat{x})\|_2 - 1)^2] \tag{2.4}$$

and the generator tries to maximize the discriminator's output for its fake instances as follows:

$$L_G = \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}[D(G(\tilde{x}))] \tag{2.5}$$

[99]: Arjovsky et al. (2017), "Wasserstein generative adversarial networks"

[100]: Gulrajani et al. (2017), "Improved training of wasserstein gans"

GANs can be extended to any kind of algorithm since they are a framework to train models in an adversarial way. However, for problems where images are used as inputs, CNNs are normally used. One of the first models to outperform other methodologies for image generation was the Deep Convolutional GAN architecture (DCGAN) presented by Radford et al. [101]. In recent years, more complex architectures have been presented, making use of different techniques that improve and overcome the GANs shortcomings, such as BigGAN [102] or StyleGAN [103]. Given these results in natural images, GANs have been applied to a variety of biomedical problems with great results (e.g. for magnetic resonance images [104], or histopathology images [105]).

[101]: Radford et al. (2015), "Unsupervised representation learning with deep convolutional generative adversarial networks"

[102]: Brock et al. (2018), "Large scale GAN training for high fidelity natural image synthesis"

[103]: Karras et al. (2021), "Alias-free generative adversarial networks"

[104]: Yu et al. (2019), "Ea-GANs: edge-aware generative adversarial networks for cross-modality MR image synthesis"

[105]: Quiros et al. (2019), "PathologyGAN: Learning deep representations of cancer tissue"

2.4 Autoencoders and Variational Autoencoders

Autoencoders [90] are another framework based on ANN for the compression of information. Similarly to GANs, they are formed by two different networks named "encoder" and "decoder". The encoder receives input and, via several hidden layers, reduces the original dimension to what is called a "bottleneck" layer. Then, the decoder uses as input this smaller representation and reconstructs it to the original input dimension. The network is trained with a mean-squared-error (MSE) loss, where the output of the decoder is expected to be as close as possible to the original input. The idea behind autoencoders is that we can learn a latent representation on this lower-dimensional representation that can be later used for downstream tasks, reducing the dimensionality of the original input. A scheme of an autoencoder can be observed in Figure 2.4.

Variational Autoencoders (VAEs) are a generalization of autoencoders that allow for better modeling of the latent space, with the aim of giving the model generative capabilities. They were firstly proposed by Kingma et al. and showed their generative capabilities on natural images [106]. In VAEs we still have the same structure as in autoencoders. They are formed by two networks, the encoder and the decoder. However, the encoder does not encode the input as a single point with the latent dimensionality, but it encodes it as a distribution over the latent space. Then, for reconstructing the input we sample

[106]: Kingma et al. (2013), "Auto-encoding variational bayes"

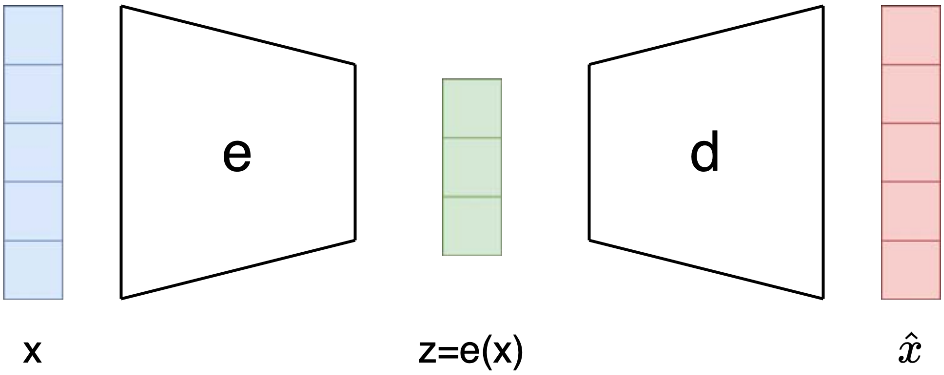


Figure 2.4: Diagram of an autoencoder, formed by an encoder and a decoder. Usually, the loss used to train an autoencoder is the mean squared error of the input (x) and the reconstruction ($\hat{x} = d(e(x))$.

from that distribution and forward it through the decoder. The assumption of the VAE is that the distribution of the data x, $P(x)$ is related to the distribution of the latent variable z, $P(z)$. Learning this distribution allows us to generate samples that, likely, will come from the distribution of the data. The loss function of the VAE, which is the negative log-likelihood with a regularizer, is as follows:

$$L_i(\theta,\phi) = -\mathbb{E}_{z\sim q_\theta(z|x_i)}[\log p_\phi(x_i|z)] + \mathbb{KL}(q_\theta(z|x_i)\|p(z)) \tag{2.6}$$

where the first term is the reconstruction loss and the second term is the Kullback-Leibler (KL) divergence between the encoder distribution $q_\theta(z|x)$ and $p(z)$ which is defined as the standard normal distribution $p(z) = N(0,1)$.

A follow-up architecture, βVAE, was later proposed with the aim of regularizing the latent space a bit more. A new parameter β, is introduced, that allows controlling the effect of the KL divergence as follows:

$$L_i(\theta,\phi) = -\mathbb{E}_{z\sim q_\theta(z|x_i)}[\log p_\phi(x_i|z)] + \beta\times\mathbb{KL}(q_\theta(z|x_i)\|p(z)) \tag{2.7}$$

If $\beta = 1$, we have the standard loss of the VAE. If $\beta = 0$, we would only focus on the reconstruction loss, approximating the model to a normal auto-encoder. For the rest of the values, we are regularizing the effect of the KL divergence on the training of the model, making the latent space smoother

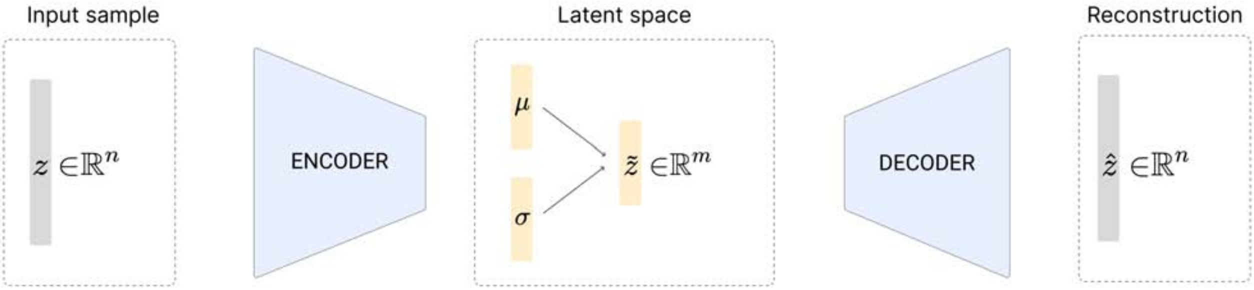


Figure 2.5: Diagram of an βVAE, formed by an encoder and a decoder. The loss combination of the reconstruction loss between the input and the output and the Kullback-Leibler (KL) divergence.

and more disentangled [107]. A diagram of a βVAE can be observed in Figure 2.5.

2.5 Support Vector Machines

KMs and more specifically Support Vector Machines (SVMs) [108, 109], are an important family of learning algorithms. They gained popularity in the mid-1990s, and since then they have been applied to multiple problems in a variety of areas with remarkable results [110–112].

In linear SVM classification, we are looking to find the maximum margin separating hyperplane where samples from different categories can be divided. The learning process automatically identifies which training samples (called support vectors) can define that hyperplane to have the wider possible gap. New samples are then mapped and predicted for each category, depending on which side of the gap they fall on. However, not all problems are linearly separable. Therefore, to loosen some of the constraints, slack variables were introduced [73]. That is, certain points will be allowed to be within the margin. We want to minimize the number of points within the margin and, subsequently, their penetration in the margin needs to be as small as possible. To this end, slack variables are introduced ξ_i, for each training data point i. Slack variables affect the optimization problem in two ways. First, they measure how much the constraint of each training data point can be violated. Second, by adding the slack variables to the energy function we are aiming to simultaneously

[108]: Cortes et al. (1995), "Support-vector networks"
[109]: Noble (2006), "What is a support vector machine?"
[110]: Lin et al. (2011), "Large-scale image classification: Fast feature extraction and SVM training"
[111]: Leslie et al. (2001), "The spectrum kernel: A string kernel for SVM protein classification"
[112]: Castillo et al. (2019), "Leukemia multiclass assessment and classification from Microarray and RNA-seq technologies integration at gene expression level"

minimize their magnitude. Another important parameter that we need to take into account is the cost function, C. The smaller this value, the stronger the regularization will be. A small C value will try to maximize the margin, being more tolerant of misclassification. Contrary, if we select a large C value, the SVM will pursue outliers more aggressively, obtaining a smaller margin but at the cost of possibly overfitting the training data [73]. Therefore, the final optimization problem can be expressed as:

$$\min_{w,b,\xi} \frac{1}{2} \parallel w \parallel^2 + C \sum_{i=1}^{m} \xi_i \tag{2.8}$$
$$\text{subject to } y_i(w \cdot x_i + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1 \ldots m$$

The success of KMs, and SVMs in particular, has been related to their effectiveness in performing a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. The kernel trick internally operates the inner product in that so-called dual space. In practice, the kernel function can be seen as a similarity measure between samples, so that when a new sample arrives, it is applied to the incoming sample with respect to the support vectors in order to select the final class of the sample. An example of how the decision margin would be defined can be observed in Figure 2.6

Although on its basis SVM is a binary classifier, some modifications can be applied to make it a multi-class classifier. The main strategy it is called One-Against-One (OVO). In the OVO classification, $K(K-1)/2$ binary classifiers are trained. At prediction time, a voting scheme is applied: all $K(K-1)/2$ classifiers are applied to a new sample, and the majority class predicted is assigned [73].

Even though DL methods have eclipsed SVMs for high-dimensional problems, they are still very competitive in medium-complexity problems, sometimes even outperforming DL techniques.

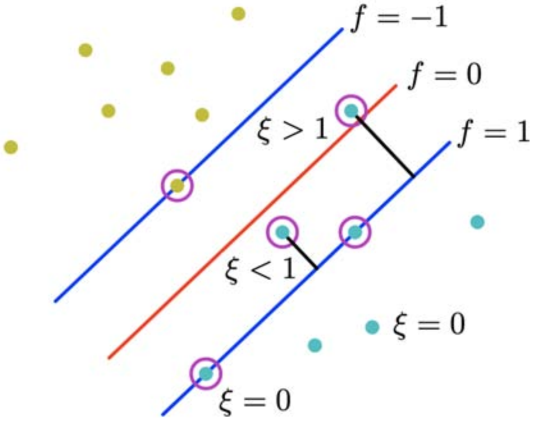


Figure 2.6: Diagram of an SVM. The decision boundary is marked by the two groups of support vectors in each class. Depending on where the new point falls, it will be classified as one or another class. Examples of slack variables can be observed within the margin and in the margin, depending on their value [73].

2.6 Information fusion in machine learning

Information fusion has been a topic of interest in ML in the last decades given the immense amount of heterogeneous information that is being gathered in problems from all areas. The main premise of these methodologies is that the fusion of the information provided by different sources can achieve better results than those obtained by independent classifiers. Three different approaches can be distinguished depending on when the fusion takes place: late, early, and intermediate fusion [113–116].

In the late fusion independent classifiers, one for each source of information, is trained over the available training data. Then, the outputs produced by these classifiers are fused in order to provide a final prediction, for instance using a weighted sum of the probabilities or by using a majority-voting scheme [117]. By doing so, the mistakes performed by some classifiers can be compensated by others, improving the final classification. In addition, using a late fusion strategy allows dealing with missing information, which is a very typical setting in biomedical problems.

In the early fusion, the features are fused before training

[113]: Daemen et al. (2009), “A kernel-based integration of genome-wide data for clinical decision support”
[114]: Gevaert et al. (2006), “Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks”
[115]: Cheerla et al. (2019), “Deep learning with multimodal representation for pancancer prognosis prediction”
[116]: Huang et al. (2020), “Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines”
[117]: Verma et al. (2014), “Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals”

the classification model [118]. This fusion can be performed through different operations, such as concatenation or summation, being the concatenation the most common one. One drawback of this approach is that it does not allow dealing with missing information in a simple way. The features of a missing modality can be set to zero, simulating the lack of information, but this can affect the training of the model.

Finally, in the intermediate fusion, the features are fused in a mid-point during the training of the classification model [115, 119]. Given these requirements, this type of fusion is more related to ANN. Thus, we would have an ANN for each data modality at the beginning, that would work as feature extractors for our data types. Then, the feature vectors obtained by forwarding these modalities are concatenated or averaged and fed to a final ANN that would perform the classification. All the ANNs involved in the architecture are trained at the same time, therefore, the classification and the feature extraction processes are linked together and optimized simultaneously.

Each of these approaches has its strengths and drawbacks. Late fusion allows having a more fine-grained control of the effect of each modality. Each modality can be, for instance, weighted for the final decision of the classifier. In addition, it can effectively deal with missing information. However, the performance of a late fusion strategy heavily relies on having strong independent classification models, and the gains obtained can be less impressive. In the early fusion, we only have one classification model, reducing the computational overhead. However, dealing with missing information is not trivial, which is an important factor in bioinformatics problems. Lastly, the intermediate fusion approach has risen in recent years as a powerful technique given the success of ANNs. ANNs have impressive feature extraction capabilities [120], which can boost the performance of the classification model. In addition, having an end-to-end training pipeline is desirable to avoid complexity. However, ANNs are known for being data-hungry, and bioinformatics problems have only a few hundred samples, limiting the use of these models. In Figure 2.7, a schematic representation of the three different options for information fusion is depicted.

Since a state-of-the-art review of these techniques is very problem-related, we have included it in each specific Chapter for the problem addressed (Chapters 4, 5, and 6).

[118]: Snoek et al. (2005), "Early versus late fusion in semantic video analysis"

[115]: Cheerla et al. (2019), "Deep learning with multimodal representation for pancancer prognosis prediction"

[119]: Cheerla et al. (2017), "MicroRNA based pan-cancer diagnosis and treatment recommendation"

[120]: Rajaraman et al. (2018), "Pretrained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images"

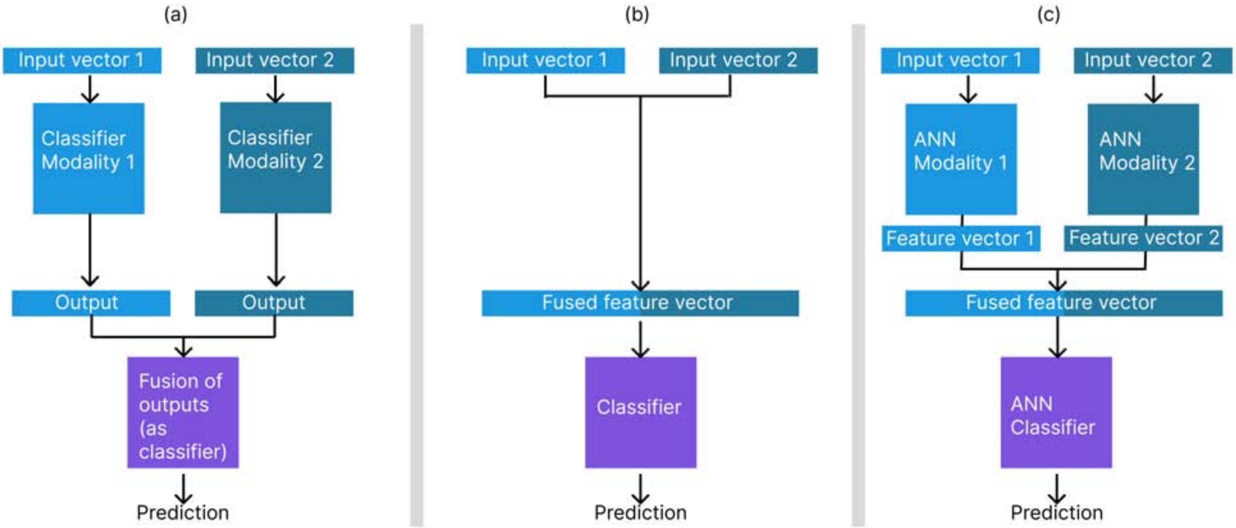


Figure 2.7: Panel A: Late fusion methodology. Two independent classifiers are trained on each data modality and their outputs are fused. Panel B: Early fusion methodology. Feature vectors for each modality are fused and a model is trained with this new representation. Panel C: Intermediate fusion methodology. The ANN feature extractor obtains a feature vector from each modality, and they are fused and input to the classification ANN. The ANNs in the architecture are trained simultaneously.

2.7 High-performance computing and accelerators in machine learning

High-performance computing (HPC) refers to the ability to perform computationally intensive operations across shared resources to achieve results in less time and at a lower cost compared to traditional computing [121]. The platform can vary, from desktop power stations to clusters of computational nodes with thousands of central processing units (CPUs). Nevertheless, the goals are the same no matter the platform: less training time, lower cost, and higher data scalability. With the increasing complexity of ML models (mainly those architectures used in deep learning problems) and the huge quantities of data that are being collected and stored now, new computing platforms have been required to reach the full potential of these techniques. For instance, CNNs (and more complex ANNs) are unfeasible to train on normal CPUs. Given their size and the number of matrix multiplications that they need to perform, other computing platforms are more adequate, like graphic processing units (GPUs) or tensor processing units (TPUs). GPUs and TPUs allow for efficient

[121]: Garcıéa-Risueño et al. (2012), “A review of High Performance Computing foundations for scientists”

parallelization of the matrix multiplications involved in both convolutional layers and fully connected layers. This allows processing more than one image at a time and also enables data scalability to huge datasets. During the development of this thesis, we used two different HPC platforms with two different implemented systems. They were used depending on the availability and the necessary requirements of the different experiments. The first one is owned by the CASIP group (ATCBIOSIMUL), from the University of Granada and the Department of Architecture and Computer Technology. The second one is part of Stanford University (Sherlock).

ATCBIOSIMUL is a computation cluster formed by one front-end node and four computation nodes, designed for different tasks. The node we used for the experiments is formed by 32 Intel(R) Xeon(R) Silver 4110 CPU @ 2.10GHz, and 128 GB of RAM memory. Using the CPUs, the different traditional ML algorithms presented in Chapters 4 and 5 were trained. The computation node is equipped with two NVIDIA GeForce TRX 2080 GPUs, with 8GB of VRAM, and they were used for the training of the CNN models presented in Chapters 4 and 5. In order to increase the number of images that could be used as input (given the small VRAM size), we decided to use the data parallel paradigm provided in the Pytorch python package [122]. With this strategy, the ANN model is replicated across multiple devices (in our case two GPUs), and the input data is evenly split between them. Once the forward pass has been performed, the metrics are averaged in one of the devices. For instance, if we are using a batch size of 4 images, two will go to GPU:0 and the other two to GPU:1, and the final loss will be computed in GPU:0. A graphical example of this strategy is presented in Figure 2.8.

The second HPC platform used during the development of this thesis is the Sherlock cluster owned by Stanford University [123], and it was used to carry out the experiments presented in Chapter 6. Sherlock is a shared computing cluster available for use by all Stanford Faculty members and their research teams, for sponsored or departmental faculty research staff. All research teams on Sherlock have access to a base set of managed computing resources, GPU-based servers, and a multi-petabyte, high-performance parallel file system for short-term storage. Sherlock features over 1,400 compute nodes, 45,000+ CPU cores, and 600+ GPUs, for a total computing power of more than 3.9 Petaflops. The cluster currently extends across 2 Infiniband fabrics (EDR,

[122]: Paszke et al. (2019), "PyTorch: An Imperative Style, High-Performance Deep Learning Library"

Table 2.1: Characteristics of the public Sherlock GPU nodes. The number of nodes, the number and type of CPUs, the amount of RAM, the computer networking communication adapter used, and the number and type of GPUs are presented [126].

Nº nodes	Nº CPUs	RAM	Computer Networking	Nº GPUs
1	20x Intel E5-2640v4	256 GB RAM	EDR IB	4x Tesla P100 PCIe
1	20x Intel E5-2640v4	256 GB RAM	EDR IB	4x Tesla P40
3	20x Intel E5-2640v4	256 GB RAM	EDR IB	4x Tesla V100_SXM2
1	24x Intel 5118	191 GB RAM	EDR IB	4x Tesla V100_SXM2
2	24x Intel 5118	191 GB RAM	EDR IB	4x Tesla V100 PCIe
16	32x AMD 7502P	256 GB RAM	HDR IB	4x Geforce RTX_2080Ti
2	32x AMD 7502P	256 GB RAM	HDR IB	4x Tesla V100S PCIe

HDR). A 5.3 PB parallel, distributed filesystem, delivering over 200 GB/s of I/O bandwidth is available. For this thesis, mainly GPUs nodes have been used, and the characteristics are described in Table 2.1. Given the size of the cluster and the number of users (more than 6,200 users), a Slurm workload manager is implemented [124]. Slurm is an open-source, fault-tolerant, and highly scalable cluster management and job scheduling system for large and small Linux clusters. As a cluster workload manager, Slurm has three key functions: allocation, managing, and contention. First, it allocates exclusive and/or non-exclusive access to resources (compute nodes) to users for a certain period of time so they can perform work. Second, it provides a framework for starting, executing, and monitoring work (normally a parallel job) on the set of allocated nodes. Finally, it arbitrates contention for resources by managing a queue of pending work. A diagram of the Slurm components is presented in Figure 2.9. Slurm consists of a $slurmd$ daemon running on each computing node and a central $slurmctld$ daemon running on a management node (as depicted in Figure 2.9. A typical workflow could be: a user prepares a bash script indicating the different requirements for their job (time, number of CPUs, number and type of GPUs, etc). The job is sent to the job queue (using *sbatch*, managed by the deamons. Then, once the resources are available, the job is put out of the queue and run in one (or more) computing node/s. Once the job has finished, or the job time has expired, the resources are freed and another job from the queue is allocated. This allows multiple users to use the same resources without a single one monopolizing them. In the Sherlock cluster, a maximum of 48 hours is set for the public computation nodes.

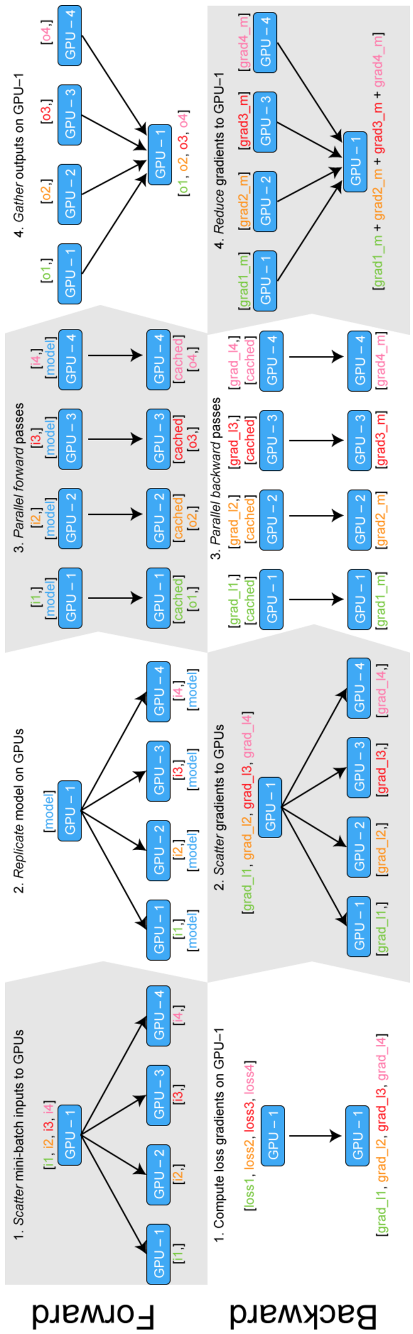


Figure 2.8: Schematic representation of the DataParallel function with four GPUs. The data is distributed across the devices and the model weights are copied to each one of them. The grouping is performed in the first GPU [125].

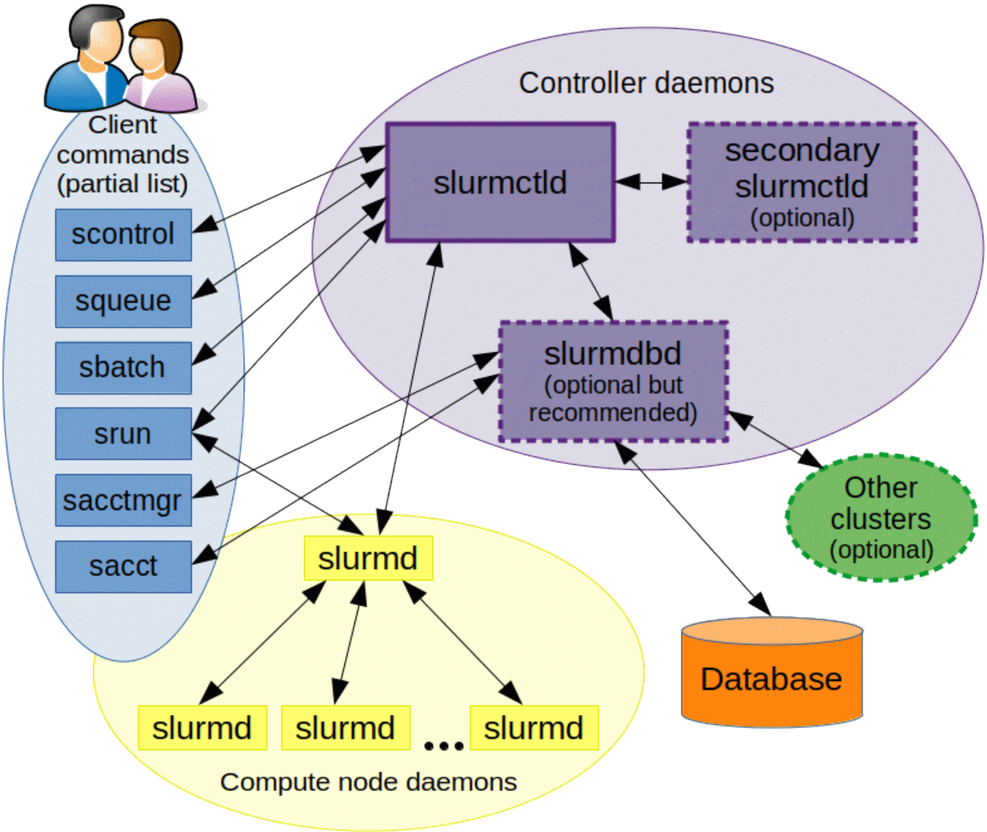


Figure 2.9: Different components that form the Slurm Workload Manager. There are commands that the clients can execute to interact with the system, and daemons that control the correct managment in the computation nodes [124].

2.8 Conclusions

In this chapter we have briefly described the main concepts in the machine learning terminology, and the most popular techniques that have been used in recent years for cancer classification and synthetic data generation. In addition, we have outlined the different possibilities available for information fusion, specifically in the context of multi-modal classification. Finally, we have described the high-performance computing platforms and frameworks that have been used along this thesis. Next, we are going to describe the different works that formed this thesis. First, we have explored the multimodal classification of NSCLC in Chapters 4 and 5, using multi-scale and multi-omics modalities. Then, in Chapter 6 we focused on the problem of multi-modal synthetic data generation, where we approached the task of RNA-to-image synthesis presenting a novel methodology for healthy image generation.

Thesis objectives

Thesis research outline | 3

After the introduction of the main concepts involved in this thesis, we are going to proceed with the research context in which this thesis was developed and the intended objectives that have been accomplished.

3.1 Research context & Motivation

As introduced in Chapters 1 and 2, the availability of multi-modal and multi-scale cancer data is rapidly increasing, allowing to use ML models for the creation of CDSS. In recent years, one of the research focus of our group was the development of intelligent ML models using and fusing heterogeneous transcriptomic data, showing the potential of these models for biomarker signature identification for cancer characterization and diagnosis support. Thanks to the collective effort of gathering multi-modal cancer data, new databases were available, allowing the exploration of information fusion for the disease and its study from a multi-scale perspective. Given the available literature at the time, we saw that there was a gap in our understanding on how these different modalities could be fused for a cancer diagnosis problem. Specifically in two of the most widely used in the current design of CDSS, such as transcriptomics data (RNA-Seq) and histopathological images (WSI). Nevertheless, there were some constraints involved. Biological data is, by default, scarce. Screening patients is very expensive, and not all tests are usually performed to a patient. Thus, the developed models needed to be resilient to missing modalities, and able to handle them. We decided to tackle the problem of lung cancer subtyping, given its importance (see Chapter 1) and the amount and balance of data available for the different modalities in the biggest multi-modal cancer database (see Chapter 2). Previous works in the literature had presented ML models for this problem, however, no multi-modal ML model was available, and no study had been perform on how different modalities interacted when using an information fusion methodology. Therefore, in the first part of this thesis, we focused on creating a ML fusion methodology for the lung cancer subtyping problem. Firstly, we approached it by using these two most well-known modalities: RNA-Seq and WSI. We decided to do so based on the prior experience the research group had with transcriptomics data, and also based on the relations between them that have been proven in literature. Once we obtained a powerful late fusion methodology, which could handle missing modalities, we concluded to firstly published those results to show how these two modalities could be used for this specific problem (Chapter 4). Next, given the importance and the need of studying the reach of the interaction among all the available omics, specially in relation to the cancer disease and the characterization of different pathological states, the study was extended to include some of those modalities available in TCGA database that could improve the performance of the fusion model. Thus, we included three new modalities (miRNA-Seq, CNV and DNA-Methylation) and studied how they interacted and which modalities worked best together. We also presented a more advanced late fusion approach, that used gradient descent to optimize the classification methodology. The results were also published in a second work (Chapter 5).

With these two works, we showed how multi-modal and multi-scale ML models can improve the performance of single-modality ML models. However, as aforementioned, not all modalities are always available. Sometimes the hospitals do not have the required equipment, or the staff is nor properly trained to perform specific tests, limiting the creation of CDSS. Synthetic data generation can help with the data scarcity problem, where data is synthetically generated with the aim to look as close to the real data as possible. However, by using only a single modality we are missing the information that other available modalities can have on the target one. Multi-modal synthetic data arise as a potential solution to this, where one modality is used to generate a synthetic paired one. Previous approaches had tried to predict the gene expression from the WSI image, but not the other way around. Therefore, and inspired by the recent advances in natural text-to-image synthesis, we focused on the problem of RNA-to-image synthesis, where the RNA-Seq profile of the patient is used to generate a synthetic WSI tile (Chapter 6). We decided to use GANs as the generative model, based on their long trajectory of successes, and infused them with the RNA-Seq profile of the patient to generate brain and lung healthy tissue, given the lower complexity in comparison to cancerous tissue. Synthetic data can be generated and used as pretraining for ML models, and then fine-tuned with the real data. In addition, it can serve for imputing missing modalities in already available datasets. There are thousands of RNA-Seq samples in the Genome Express Omnibus (GEO) for which the paired WSI is not available. By using our presented models, a corresponding synthetic set of WSI tiles can be generated using the RNA-Seq profile of the patient, allowing the training of ML models in CDSS for cancer.

3.2 Objectives

Based on the aforementioned motivation based on the research literature of multi-modal ML CDSS for cancer, the objectives tackled in this thesis are the following:

1. **Develop new ML methodologies for the fusion of heterogeneous biomedical data for cancer problems:** new ML models should be developed for multi-modal and multi-scale cancer modalities, using the state-of-the-art preprocessing, analysis and data processing steps for each data modality used. Besides, specific attention has to be given to the study and development of novel ML fusion methodologies for the multi-omics problems approached. Two specific problems will be tackled:
   a) Study the data fusion possibilities of the two most independently used modalities in CDSS in cancer research such as WSI and RNA-Seq, specifically for NSCLC.
   b) Study the limits of adding more heterogeneous biomedical data sources, including miRNA-Seq, DNA Methylation and CNV data modalities, also for NSCLC.
2. **Develop new ML methodologies for the generation of multi-modal biomedical synthetic data to fight data scarcity:** develop a multi-modal synthetic data generator using ML methodologies, able to generate one modality based on a paired one.The proposed methodology should be evaluated both in-silico and against expert pathologists.

Secondary objectives were also tackled during the development of this thesis:

1. **Analyze the robustness of the proposed methodologies under missing or incomplete information:** the proposed ML methodologies for the fusion of biological information must be robust against missing modalities. They need to provide a prediction even when some modalities are missing.
2. **Obtain a robust and reliable set of cancer biomarkers for the different data modalities used:** state-of-the-art preprocessing techniques should be used to obtain a set of biomarkers, adjusting them to each modality. Special attention will be provided to achieve biologically relevant genes from the RNA-Seq dataset for NSCLC classification. The obtained biomarker set used for the classification purposes must be robust and reliable.
3. **Optimize of the proposed methodology using high-performance computing architectures:** adequate high-performance computing platforms (such as clusters of GPUs) must be used for the training of the different modalities, according to the different requirements that these present.

Research articles

4 Non-small-cell lung cancer multi-modal classification using RNA-Seq and WSI

Source

Francisco Carrillo-Perez, Juan Carlos Morales, Daniel Castillo-Secilla, Yésica Molina-Castro, Alberto Guillén, Ignacio Rojas, Luis Javier Herrera. "Non-small-cell lung cancer classifi- cation via RNA-Seq and histology imaging probability fusion". In: *BMC bioinformatics* 22.1 (2021), pp. 1–19.

Abstract

Adenocarcinoma and squamous cell carcinoma are the most important subtypes of non-small-cell lung cancer (also known as lung carcinoma), the most common lung cancer type. Their distinction is crucial in order to provide an accurate treatment to the patient and thus enhance their prognosis. Multiple screenings are normally carried out to detect the disease, including visual evaluation of histology slides by an expert or the analysis of the gene expression of the patient to find cancer-driven genes. However, given the limited time available to clinicians, the creation of clinical decision support systems that can automatically extract diagnostic information from these biological sources has rapidly increased (e.g. from histology imaging, next-generation sequencing technologies data, clinical information, etc.). These systems usually are fed a single data modality, leaving out the multi-scale and multi-omic nature of cancer data. Besides, by fusing the information provided, we can mimic the interaction between doctors in a hospital when they provide a final diagnosis. In this work, we present a late fusion classification model using whole-slide images and RNA-Seq data for the diagnosis of non-small-cell lung cancer, classifying among its two most common subtypes and control patients, i.e. adenocarcinoma, squamous cell carcinoma, and control patients. The late fusion model improves the performance of independent classifiers, reducing the diagnosis error rate up to a 64% over the whole-slide-imaging classifier and a 24% over the isolate RNA-Seq classifier, and obtaining a mean F1-Score of 95.19% and a mean AUC of 0.991. The obtained results highlight that

the fusion of information can improve the performance of single-modality classifiers by using the interaction between the biological data in cancer and alike diseases, where a multi-modality and multi-scale nature exists.

4.1 Introduction

Appropriate identification of the NSCLC lung cancer subtype is critical in the diagnostic process since therapies differ for LUAD and LUSC [127] (see Chapter 1). Therefore, finding accurate and robust biomarkers in different types of patients' biomedical information is crucial to accelerate this process. Experts use several methods for lung cancer type classification, such as Computer Tomography screening, WSIs, the identification of biomarkers in next-generation sequencing (NGS) data (e.g. gene expression analysis using RNA-Seq), or the use of clinical information from the patient. The manual analysis of these sources of information can be a time-consuming and exhausting task. Thus, in recent years the automatic analysis of each of the aforementioned data types has been explored [128–132]. However, the majority of the proposed models are single-modality classifiers, limiting the potential that can be obtained by fusing the data sources. This is especially interesting in the case of cancer, where biological data sources have a strong relation between them. For instance, mutations can be found in the genome, showing effects of the tumor microenvironment and visually modifying the tissue [133–135]. Therefore, by not having all modalities available, we are losing a part of the picture that can lead to early detection of the disease.

Using information fusion methodologies that integrate the predictions of systems using biological information may enhance the diagnosis of a patient. Information fusion has been a topic of interest in ML research in late years based on the growth of multi-modal data. The fusion of biological data has been mainly explored in literature for prognosis prediction, obtaining good results [115, 136–142]. Since not all sources of information are always available, having an integration model for the classification would also allow predicting the lung cancer type even when only one source of information is available. A model of these characteristics would fall into the design of decision-making support systems that are applied to precision medicine [143], as the immediate future of bioinformatics and medicine. To the best of our knowledge, the integration of gene expression data and biomedical imaging to provide a classification model for LUAD, LUSC, and control patients has not been proposed in the literature.

The aim of this work is to present a classification model using a late fusion methodology for the task of LUAD, LUSC, and control patients classification by fusing the probabilities obtained by two classifiers. One classifier uses as input RNA-Seq data and the other one WSI. In this Section, an introduction to the problem has been outlined. In Section 4.2, an overview of the related works in the area of ML applied to lung cancer will be reported. In the Section 4.3, the methodology and data used for this work will be presented. In the Section 4.4, results obtained for the proposed experiments will be shown and discussed. Finally, in Section 4.5 conclusions will be drawn and future work will be outlined.

4.2 Related work

Lung cancer gene expression classification

Over the last few years, the potential of ML models using NGS data for cancer classification has been shown. Specifically, several works can be found in literature for lung cancer type classification using gene expression data.

Since LUAD is the most frequent lung cancer type, many works have been published for LUAD and control classification. Smolander et al. presented a deep learning model using gene expression from coding RNA, and non-coding RNA [144]. They obtained a classification accuracy of 95.97% using coding RNA. Similarly, Fan et al. using Support Vector Machines (SVMs) and a signature of 12 genes reached an accuracy of 91% [145]. Zhao et al. combined the information from ncRNAs, miRNAs and mRNAs for the classification, using SVMs. Finally, they selected 44 genes and they reached a classification accuracy of 95.3% [146] .

For lung cancer subtypes classification, Gonzales et al. studied differentially expression genes (DEGs) between SCLC, LUAD, LUSC and Large Cell Lung Carcinoma. Then, different feature selectors and predictive models were used in order to compare their classification performance [147]. Authors reached an accuracy of 88.23% using k-NN and the Random Forest feature selector.

Lung cancer histology imaging classification

In recent years, the use of deep learning models for histology imaging classification has been taken into consideration based on the outstanding results that these models have reached in computer vision tasks and in health informatics [88, 148]. Deep learning models require huge quantities of data in order to properly learn features from an image. Therefore, the most popular approach for histology image classification has been to perform a segmentation of regions of interest of each slide (or placing a label for the whole slide). Then tiles can be extracted and labeled by experts from the image for a posterior training. Thus, a huge increment in the available dataset is obtained.

Based on the aforementioned methodology, different works have been published for lung cancer classification. Coudray et al. used a deep learning model using transfer learning for LUAD, LUSC and control classification and mutation prediction, reaching a 0.978 score of Area Under the Curve (AUC) [131]. Tiles were extracted and were used to perform the training and the classification. Similarly, Kanavati et al. presented a deep learning model using transfer learning for lung carcinoma classification, reaching a score of 0.988 AUC for a binary classification problem [149]. Authors used labeled images by experts for their work. Graham et al. presented a two steps methodology for LUAD, LUSC and normal classification using a deep learning model trained on image tiles and then extracting summary statistics from them for the final classification, obtaining an accuracy value of 81% [150]. Likewise, Yi et al. trained and compared the performance of several CNNs with different structures in classifying image tiles as malignant vs. non-malignant, obtaining an AUC score of 0.9119 [151].

Yu et al. combined traditional thresholding and image processing techniques for slides images with machine-learning methods, achieving an AUC of 0.85 in distinguishing normal from tumor, and 0.75 in distinguishing LUAD from LUSC [152]. Khosravi et al. used deep learning for the classification of breast, bladder and lung tumors, achieving an AUC of 0.83 in classification of lung tumor types on tumor slides [153].

Fusion of omics and histology data

When it comes to the integration of information from different omics data and histology imaging, different approaches have been proposed in the recent literature, mainly for prognosis prediction.

Lee et al. presented a multi-modal longitudinal data integration framework based on deep learning to predict Alzheimer disease progression [138]. In this case, MRI scans, genomics information and cognitive assessments were used as inputs. In order to obtain the feature representation, a recurrent neural network with gated recurrent units [139] was used. Then, features were concatenated and a final prediction was performed.

Another methodology that has recently been presented in literature with remarkable results is obtaining a feature vector for each type of data and then performing a plain concatenation of those features vectors or applying attention mechanism before the concatenation for model training. Lai et al. developed a deep neural network (DNN) combining heterogeneous data sources of gene expression and clinical data to accurately predict the overall survival of NSCLC patients [140]. The combination of 15 biomarkers along with clinical data were used to develop an integrative DNN via bimodal learning to predict the 5-year survival status of NSCLC patients with high accuracy. The combination outperform the results obtained for each type of data separately. Silva et al. presented an end-to-end multi-modal Deep Learning method, named MultiSurv, for automatic patient risk prediction for a large group of 33 cancer types [141]. They compared fusing different sources of information by using attention weights, and used a feed-forward neural network for predicting. Chen et al. presented an integrated framework for fusing histology and genomics features for survival outcome prediction [142]. The authors used WSIs, mutations information, Copy Number Variation (CNV) information and mRNAseq. Different features were obtained and fed to an attention mechanism, that was later used for survival prediction and grading. The results presented by the authors shown that the use of the fusion outperform the results obtained by using each type of data separately.

Finally, several works have been published where features used for classification have been obtained using autoencoders [82]. Cheerla et al. presented a deep learning model with multi-modal representation for pancancer prognosis prediction [115]. The survival prediction of patients for 20 different cancer types was performed using as information clinical data, mRNA expression data, miRNA expression data and WSIs. Feature vectors were obtained and then combined for prognosis prediction. They obtain outstanding results, specially in those cancer types where not many samples were available. Simidjievski et al. investigated several autoencoder architectures to integrate a variety of patient data such as gene expression, copy number alterations and clinical data, showing the usefulness of this approach for different breast-cancer analysis tasks [136]. Following with the use of autoencoder architectures for information integration, Ma et al. proposed a Multi-view Factorization AutoEncoder (MAE) which not only encodes gene expression, miRNA expression, protein expression, DNA Methylation and clinical information but also includes domain knowledge such as molecular integration networks for bladder urothelial carcinoma and brain lower grade glioma classification [137].

4.3 Methods

Data acquisition

In this work, we have used two different types of biological data: RNA-Seq and WSIs. The data were gathered from The Cancer Genome Atlas (TCGA) program [57], located in the GDC portal [154].

The TCGA contains information from 33 different cancer types, and they have achieved the goal of providing easy access to the data. In addition, GDC have performed an harmonization of all the available samples in the program. Moreover, various data types are available for each sample (e.g. gene expression, copy number variation, histology imaging, etc.). In GDC, each patient has a Patient ID that identifies them, and each screening performed on the same biological sample from a patient has a defined Case ID. Therefore, for each Case ID we can have different biological information (WSI, RNA-Seq or both). Those Case IDs used in this work are available in this Github repository *. Table 4.1 shows the Case IDs availability per class and considered data type.

Table 4.1: Number of samples per class for each data type.

	WSI	RNA-Seq	Both
LUAD	495	457	442
Healthy	419	44	41
LUSC	506	479	467
Total	1420	980	950

WSIs data needs to be preprocessed prior to any analysis. The preprocessing of WSIs relied on the Python package openslide [155], that efficiently read and parse these type of images.

For the case of gene expression, RNA-Seq from Illumina HTSeq data is used in TCGA. In the specific case of GDC data, it harmonizes RNA-Seq data by aligning raw RNA reads to the GRCh38 reference genome building and calculating gene expression levels with standardized protocols [156]. The KnowSeq R-Bioc package was used in order to obtain the values of the DEGs [157].

Models were implemented in Python with the Pytorch [122] and Scikit-Learn [158] packages. Deep Learning model training was performed using a NVIDIA^{TM} RTX 2080 Super GPU.

In order to avoid a result bias due to a reduced test set and the data imbalance, the whole dataset was divided using a 10-Fold Cross-Validation (10-Fold CV), in order to obtain a more thorough assessment of the proposed methodology. In each iteration of the 10-Fold CV process, the training set was used to train the models, and also for hyperparameter tuning, while a final assessment of models performance was done in the test set. The hyperparameter tuning strategy used differs for each data type and will be later explained in each model section; concretely, a single traditional training-validation subdivision was used for the WSI model, and grid search CV was used for the RNA-Seq model.

All the splits were performed both in a patient-wise way and in a stratified way. With a patient-wise splitting we are

* https://github.com/pacocp/NSCLC-multimodal-classification

Table 4.2: Number of tiles per class.

	# Tiles
LUAD	100,841
Healthy	62,715
LUSC	92,584
Total	256,140

ensuring that, even if a patient has more than one case, they could only belong to one of the splits, being this training or validation. Imposing this restriction prevents any kind of information leakage during training. On the other hand, through stratified splitting the proportion of classes in each fold is maintained.

WSI preprocessing

WSIs, also known as virtual microscopy, refers to scanning a complete microscope slide and creating a single high-resolution digital file. With it, different resolutions of the same image can be obtained and an extraction of tiles can be performed. The generated file has SVS format, and several preprocessing steps need to be taken in order to work with this type of files. Firstly, SVS images are read with an specific factor of magnification. In this work, a factor of 20x was chosen in order to obtain an adequate resolution (leaving images with an approximate resolution of 10,000x10,000). Once images were obtained, we converted them from SVS to PNG format in order to facilitate further manipulations.

We took non-overlapping tiles of 512x512 omitting those tiles where most of it was white background. To test this condition, we computed the mean value for each channel. If in all three channels the value was greater than 220, the tile was discarded, otherwise it was selected as proposed in literature by Coudray et al. [131]. This methodology allows to multiply the number of images available to train the model, since we are using all the tiles that can be extracted from each one of the images instead of only using the whole image. This enables the deep learning models to more easily learn relevant features for the classification task. The number of tiles per class is depicted in Table 4.2.

RNA-Seq data preprocessing

In order to analyze the HTSeq-Counts data provided, we used the KnowSeq R-Bioc package [157]. This package provides a pipeline to obtain DEGs based on the HTSeq-Counts files and then performs a machine learning assessment of the selected DEGs. KnowSeq relies on limma [159], which is the state-of-the-art for finding differential expressions. However, limma is usually employed to biclass problems, where two classes need to be compared. Thus, additional tasks need to be perform to achieve DEGs when there are more than two classes. In order to deal with this problem, Castillo et al. presented the coverage (COV) parameter, which uses limma to a perform binary comparisons of the N presented classes and finally select a set of genes that are differentially expressed in COV binary comparisons [112].

Therefore, we used the *DEGsExtraction* function from the KnowSeq package over the training set for obtaining the DEGs matrix. As parameters, a Log_2 *Fold Chain* (LFC) value of 2, a p-value of 0.05 and a COV value of 2 were set. Once we obtained the DEGs matrix, we used the minimun Redundancy Maximum Relevance (mRMR) algorithm to obtain a ranking of the genes [160]. The mRMR algorithm uses information theory for obtaining a ranking of features which are highly correlated with the classes but with a minimum redundancy between them. mRMR algorithm has been previously used in literature as feature selector for gene expression [112, 161–163].

Database organization

To facilitate the handling of data, we created a local database structure that organizes the data per patient and sample. Since the system is oriented toward helping doctors to predict cancer diseases, the database needs to be easily scalable and accessible. This means that it needs to be human-readable, as well as to have a standardized structure. The last part is essential, since it allows for more data to be added to the database, and it facilitates the task of creating the different classifiers. Therefore, this work presents a database structure that fits these requirements. In Figure 4.1, a simplified representation of the database for a single patient can be observed.

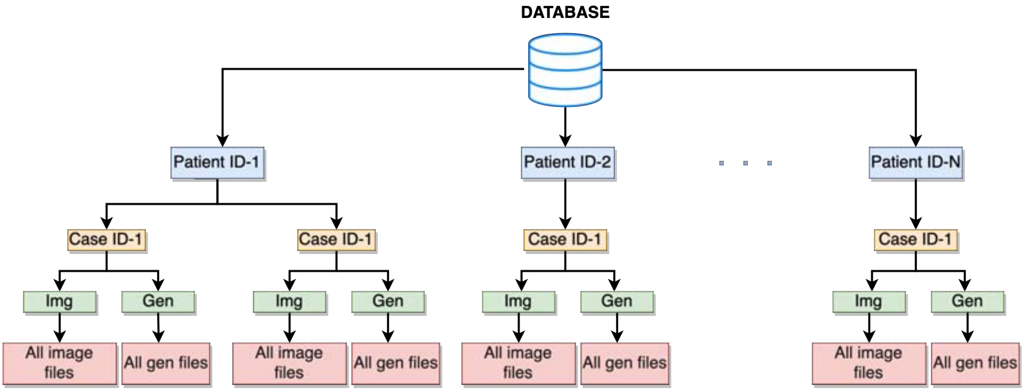


Figure 4.1: Schematic diagram of the database organization for a given patient. Any other type of biological information can be added to the database. Simply, a new folder with the new biological information needs to be added to the given Case ID folder of the patient.

The database follows the structure "patient → case → data type". The root of the database contains multiple directories inside, where each of which is associated with a single patient. There is no directory that contains data from two different patients, nor two directories that contain information about a single patient. Inside each patient directory, there are many more directories, each one corresponding to a case. All the information of a case is stored in the corresponding case directory. Once again, inside the case directory, there are other directories, each corresponding to a specific data type such as gene expression data or histology images, and where other types of information can be easily added. All the directories that correspond to the same data type need to maintain the same naming convention among the different cases and patients for scalability. Given the case that a certain data type does not exist for a particular case directory, its corresponding data type directory won't exist. This is especially relevant since it implies that the model needs to behave well even in the absence of specific information.

Per-tile model and per-slide classification

For the per-tile classification we used a CNN, given that they have proven to be the state-of-the-art in computer vision problems, using a transfer learning approach. CNNs have been widely used in literature for WSIs classification with great results, as described in subsection 4.2. In addition, given the size of the tile dataset, using other classical machine learning models might not be computationally feasible. Transfer learning allows to use the filters that have been learned in another problem domain with sufficient data, and adjust the weights of the network to another given problem. Different architectures were tested such as VGG-16 [164] or Efficientnet [165]. Finally, the Resnet-18 architecture [166] was used with pre-trained weights on Imagenet [167], and tiles were normalized with the mean and standard deviation from it. The classification layer of the architecture was adapted to the set of classes, but preserving the same structure. Only the last residual module of the architecture was fine-tuned, the rest of the weights of the network were frozen. As it is usual for deep learning models, in each split a 10% of the training data was used as validation set for the network optimization and hyperparameter tuning.

For the CNN training 25 epochs were used, monitoring the accuracy on the validation subset with the early stopping methodology and saving the best weights for later use. As a loss function, the cross entropy loss function was selected. As the optimizer, Adam was chosen with a learning rate (LR) value of $1e^{-5}$, betas equal to $(0.9, 0.999)$ and epsilon equal to $1e^{-8}$. Since the Resnet-18 is being fine-tuned, a small LR needs to be used or the pre-trained weights will change more than desired. These hyperparameters were chosen by manually tuning them during the experimentation, and based on results in the validation set.

Once we have obtained a per-tile model, we now need to define how to classify a slide. In this work, we used a majority voting approach, similar to the methodology presented by Coudray et al. [131]. Having all tiles from the image classified using the per-tile model, the most predicted class among all the tiles is used as the final prediction. Variations using different thresholds (instead of simple majority voting) to choose the final prediction were inspected. Nevertheless, in our case, the aforementioned methodology provided the best performance. The training time of the model was around 14 hours using the ATCBIOSIMUL cluster presented in Chapter 2.

RNA-Seq classification model

For the RNA-Seq feature extraction, we carried out the preprocessing steps explained in subsection 4. Selecting how many genes to perform the classification with is of utter importance since usually, clinicians are looking for the smallest gene signature that led to good classification performance. This decision is important due to the necessity of providing a small gene signature that can facilitate its use in a standard clinical laboratory, for instance in a PCR-based diagnosis assay [168–170]. We used three different sets of genes (3, 6, and 10) in order to compare the performance of the fusion model when using comparatively small, medium, and large size of gene signatures.

It is important to note, that the simulations performed under the 10-Fold CV assessment, it implies using different training datasets for the gene signatures extraction. This could lead to small variations in the signatures obtained [171] since we are using a different group of samples as a training set each time. The final gene signature proposed was the one formed by those genes that are best ranked by the mRMR algorithm in the rest of the gene signatures.

We tested different classification techniques for the RNA-Seq classification task such as K-Nearest Neighbors, SVMs, or Random Forest. Finally, SVMs were selected, since they outperformed the rest of the models in the validation set and they have proven to be really successful in mid-size problems. In addition, it has been used in the gene expression literature for cancer classification with good results [112, 161, 162]. A grid search CV was used over the training set in each split for the parameters optimization, using the Gaussian Radial Basis Function kernel as it has proven to offer a good asymptotic behavior [172]. The search range of values for both C and γ was: $[2^{-7}, 2^{-5}, 2^{-2}, 2, 2^{4}, 2^{7}]$. Moreover, features were normalized between -1 and 1. The training time of the model was around 30 minutes using the ATCBIOSIMUL cluster presented in Chapter 2.

Probability Fusion

As it has been reported in the Section 4.1, among the different approaches proposed in the literature for data fusion, this work uses a late fusion methodology. We also performed experiments using early fusion approaches, in which obtained features from both RNA-Seq and WSIs data types were concatenated and fed to a classifier performing the final prediction. Under this last scheme, the straightforward features extracted for each data type (gene expression on one side, and accumulation -average sum- of the features extracted from the CNN for the different tiles of an image) were observed to decrease the performance of the fusion classification model. This decrease in the performance in comparison to the late fusion model may be due to the difference between the dimensionality of the features obtained from each data type since a feature vector of size 512 is obtained in the case of the WSI and a feature vector of size between 3 and 10 genes is obtained in the case of RNA-Seq. Even after applying a number of approaches to reduce the feature dimensionality, such as PCA or max pooling operation on the CNN features, results were not surpassing the late fusion scheme next explained.

There exist two options when applying a late fusion approach: to combine the predictions or to combine the probabilities returned by each classifier. These two approaches rely on the classification models used. The first option, i.e. integrating the predictions, would require weighting the classifiers, since only two data types are used. The second, the fusion of the probabilities, allows including more information for the classification (i.e. the probabilities assigned by each classifier to each class). So this last option was selected in the expectancy of a more powerful information fusion.

The probabilities for each classifier were obtained as follows. For the RNA-Seq model, the probabilities are returned by the SVM by using the methodology proposed by Wu et. al. [173], which uses a pairwise coupling method that is included in the Scikit-Learn library [158]. With this methodology, we are able to obtain three probabilities, one per class, which model the belongingness of a sample to each class, and where the sum of the probabilities is equal to one. On the other hand, for the WSI we need to compute them. In the per-slide classification, we have the number of tiles predicted per class, therefore, we compute the probability per class for each slide as the number of tiles predicted for a class divided by the total number of tiles in the slide (see Eq. 4.1).

$$P^{CNN}(x, c_i) = \frac{\#TilesPredicted_{(x,c_i)}}{\#SlideTiles} \tag{4.1}$$

where x is the sample to be predicted and c_i is the given class: LUAD, Healthy, or LUSC. Using this methodology we are able to obtain three probabilities (one for each class) representing how likely is for that slide to belong to each one of the classes, depending on the predictions provided by CNN. In addition, and given that the number of predicted tiles per class is divided by the total number of tiles, the sum of the obtained probabilities is equal to one.

Once we have obtained the probabilities for each classifier and class, we need to fuse them to make a final prediction. In this work, we propose a weighted sum of the two probabilities by using two weight parameters: α_1 and α_2 (see Eq. 4.4). They will control the trade-off between the probabilities returned by the two models: α_1 for the WSI CNN classifier (P^{CNN}) and α_2 for the RNA-Seq SVM classifier (P^{SVM}). This will allow both classifiers to support each other's predictions: in case one of the classifiers is providing a borderline wrong decision, the other one could balance it to the right side.

Some approaches have been proposed in the literature to weight the probabilities obtained from classifiers using different modalities. Dong et al. proposed to give a weight to each classifier based on the performance of each model [174]. Similarly, Meng et al. proposed to compute the weight based on the accuracy of each model applying a normalization between the maximum and the minimum accuracy [175]. Trong et al. proposed to normalize the accuracy only based on the maximum accuracy achieved in order to obtain a weight [176]. Other approaches have been taken, such that proposed by Depeursinge et al. where the probabilities returned by two SVMs were multiplied and the maximum was chosen for the prediction [177].

In this work, the weight for each classifier is computed based on their mean performance in ten different stratified resampling sets obtained from the training set. Resampling is a methodology that consists of taking a random subset of samples from a given set, usually a percentage of it, and has shown to be useful for robust statistic estimation [178]. In this work, a 90% of the training set was randomly chosen for each resampling in a stratified way, i.e. maintaining the percentage of each class in each set. Due to the imbalance of the dataset, the F1-Score metric was chosen as the performance measure. Thus, firstly the mean of the F1-Score metric is obtained across the ten different resamplings of the training set as follows:

$$\overline{F1}_M = \frac{\sum_{i=1}^{10} F1_{M_i}}{10} \tag{4.2}$$

where $\overline{F1}_M$ is the mean F1-Score of a model M (i.e., CNN or SVM) across all the resampling sets and $F1_{M_i}$ is the F1-Score metric obtained by the model M in the ith resampling set.

Then, after computing the mean of the F1-Score for each model across the 10 different resampling sets, the final weight for each classifier is computed as follows:

$$\alpha_1 = \frac{\overline{F1}_{CNN}}{\overline{F1}_{CNN} + \overline{F1}_{SVM}} \alpha_2 = \frac{\overline{F1}_{SVM}}{\overline{F1}_{CNN} + \overline{F1}_{SVM}} \tag{4.3}$$

where $\overline{F1}_{CNN}$ and $\overline{F1}_{SVM}$ are the mean F1-Score obtained by each model across the resampling sets, and it is satisfied that $\alpha_1 + \alpha 2 = 1$.

These α_1 and α_2 are then used for the fusion model in order to weight the probability returned by each classifier. Thus, the probability for a sample x belonging to a class c_i will be calculated using the following equation:

$$P^{^{Fusion}}(x, c_i) = \alpha_1 * P^{^{CNN}}(x, c_i) + \alpha_2 * P^{^{SVM}}(x, c_i) \tag{4.4}$$

With this automatic methodology, we are getting a weight value for the classification that allows fusing the classifiers' probabilities based on how well they performed on the training set. Given that we are using a 10-Fold CV for evaluating the methodology, different α values might be obtained for each split, since different training sets are being used. It is important to note that this fusion methodology allows us to effectively deal with missing information. If one of the data types is missing, then only the probabilities of the other classifier will be taken into account, without the need to average them. The pipeline for estimating the probability of a given sample belonging to the class is depicted in Figure 4.2

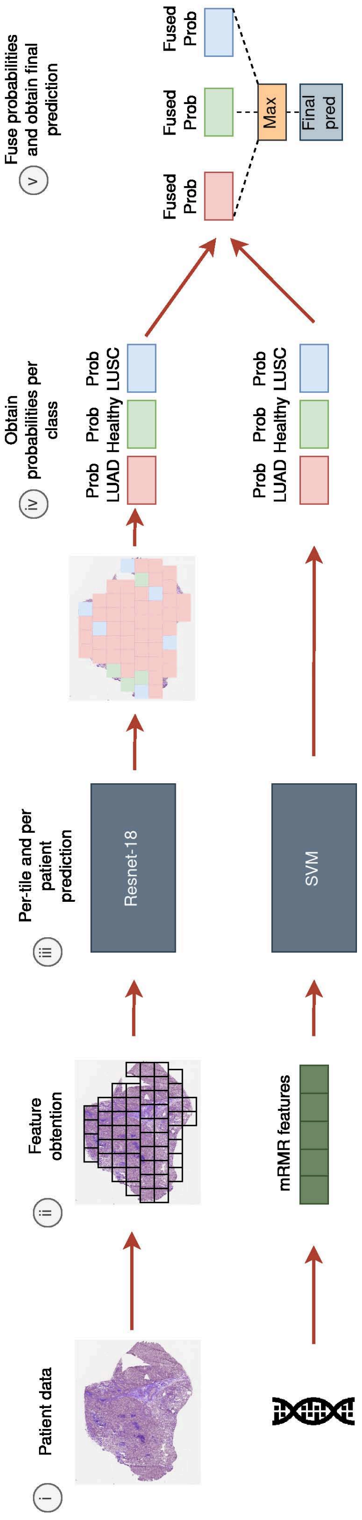


Figure 4.2: Pipeline for sample prediction. (i) Both the WSI and the RNA-Seq data for that case ID are obtained. (ii) Non-overlapping 512x512 tiles are extracted for the WSI, filtering the background. For RNA-Seq, we took the set of DEGs selected by the mRMR ranking. (iii) For the WSI, the probabilities are obtained by averaging the number of tiles predicted per class and the total number of them. For RNA-Seq data, the probabilities are returned by the SVM. (iv) We fuse the probabilities by averaging the ones obtained by each classifier per class, and the final prediction is the class with the higher probability.

4.4 Results and Discussion

Classification performance of the models

The presented results are for those cases where both sources of information are available (see Table 4.1), allowing a fair comparison of the improvement that can be obtained under the information fusion approach. Models were trained on all the available data in each training set, and a global assessment is presented using a 10-Fold CV approach on the whole dataset. All results are presented for the CNN per-slide classification using WSI as input data, the SVM using RNA-Seq data, and the fusion model. We computed the accuracy, F1-Score, confusion matrices, ROC curve, and Area Under the Curve (AUC) for each set and split.

Table 4.4 shows the accuracy, F1-Score, and AUC for the WSI classifier, the RNA-Seq classifier, and the fusion model. Results are averaged for the ten executions and the standard deviation obtained is also shown.

With respect to the RNA-Seq classifier, three different set configurations were tested: 3, 6, and 10 genes. A comparison of using different sets of genes for the fusion and RNA-Seq model can be observed in Figure 4.3. The RNA-Seq model obtains good results across the splits for the three configurations, with relevant improvement observed when using 6 genes over 3. However, a very similar performance is observed when using 10 genes (94.05% of F1-Score, and 94.12% of accuracy) in comparison with 6 (93.67% of F1-Score, and 93.70% of accuracy), even with a higher standard deviation when using 10 genes. This enables to choose a gene expression model with 6 genes without significantly affecting the performance in comparison to using a larger gene set, which facilitates its utilization in a standard clinical laboratory [168, 169]. For the AUC metric, the model also achieves impressive results with both sizes, reaching 0.987 and 0.990 respectively. These results are comparable to those obtained for a binary classification problem (LUAD vs Healthy) where authors reached an accuracy of 95.97% and 91% (see subsection 4.2).

In relation to the WSI data classification, Table 4.4 shows that this model presents a lower classification performance in comparison to the RNA-Seq and fusion model, achieving an F1-score of 83.39% and an accuracy of 86.03%. For the AUC metric, the results are similar or even improve in some cases those obtained in literature (see subsection 4.2), achieving 0.947 in the validation set across the splits.

The fusion model was optimized using the methodology proposed in the subsection 4, choosing an optimized value of α_1 and α_2 for each split. It must be noted that the range of the α values obtained in each configuration was very similar across the splits, with α_1 ranging from $[0.49 - 0.52]$ and α_2 ranging from $[0.48 - 0.51]$. The fusion model outperforms the RNA-Seq and WSI models for all metrics (see Table 4.4). Also, no matter the number of selected genes, the fusion model always outperforms the gene expression model (see Figure 4.3). The configuration of the fusion model using 3 genes is slightly outperformed by the RNA-Seq model using 6 genes. This is due to the lower performance of the RNA-Seq model configuration when using 3 genes. However, the fusion model still achieves a better performance in comparison to the WSI and the RNA-Seq configuration using 3 genes (see Table 4.4). Taking the configuration with 6 genes, the fusion model achieves a mean F1-Score of 95.19%, a mean AUC of 0.991, and a mean accuracy of 95.18%. For that model, Figures 4.5 and 4.6 show the confusion matrices and the ROC curves for the whole dataset. For the fusion model, similar results are obtained when using 6 and 10 genes, which allows using the model with a smaller gene signature. Given the low number of healthy samples where both data types are available (see Table 4.1), it is interesting to note that the mean F1-Score achieved is high, which means that these are being correctly classified on the whole dataset. The standard deviation of the metrics across the splits decreases with the fusion model, showing that it allows a more stable behavior than the separate SVM and CNN models. The results obtained in the classification problem are also comparable to those obtained in literature, reaching those accuracies obtained in a binary classification problem when using RNA-Seq data (95.97 % [144], 91% [145], 95.3% [146]) and the AUC obtained when using WSIs as input for the multi-class classification (AUC 0.978 [131]) and for the binary classification (AUC 0.988 [149]).

In order to visualize the performance of each model per class, we plotted the ROC curves for the whole dataset (see Figure 4.6), for the fusion model with 6 genes. Confusion matrix was also extracted for the whole dataset (see Figure 4.5). As it can be observed, the fusion of probabilities obtains a better

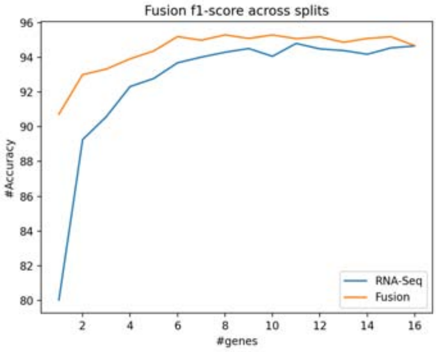


Figure 4.3: Performance of the fusion or RNA-Seq model in terms of F1-Score depending on the number of selected genes. The fusion model always outperforms the RNA-seq model.

performance for the three classes over the CNN and the SVM models. In addition, the fusion model reduces the number of missclassified samples from 133 and 60 to 46, for the CNN and the SVM respectively, over the whole dataset when using 6 genes (see Table 4.5). This represents an improvement of the error rate of ≈ 65% over the CNN, and ≈ 24% over the RNA-Seq model.

Based on the results we have obtained, the fusion model is correctly classifying samples that one of the models was wrongly predicting (see Figure 4.5 and Table 4.5). We analyzed the models predictions for the whole dataset to assess the cases in which both classifiers were providing different outcomes. An example can be observed in Figure 4.4.

Finally, in order to provide a biologically relevant single gene signature for clinical use, the use of a single gene signature was inspected. As the final unique gene signature, we selected the one from the ten obtained in the 10-Fold CV process whose genes appeared in the first positions of the mRMR ranking for the rest of the splits. The 6-genes signature is formed by the following genes: *SLC2A1,NTRK2,TOX3,NXPH4,TFAP2A,KRT13*. The correlation of these genes with lung cancer was verified in the Open Targets platform [179], whose association scores with cancer, lung cancer, NSCLC, LUAD and LUSC, are shown in Table 4.3. Its performance was evaluated over the 10-CV, achieving a mean F1-Score, AUC and accuracy of 94.35%, 0.985 and 94.32% respectively for the isolate RNA-Seq SVM model, and 95.31%, 0.991 and 95.29% for the fusion model. In addition, a biological relevance analysis of these DEGs can be found in Section 4.4.

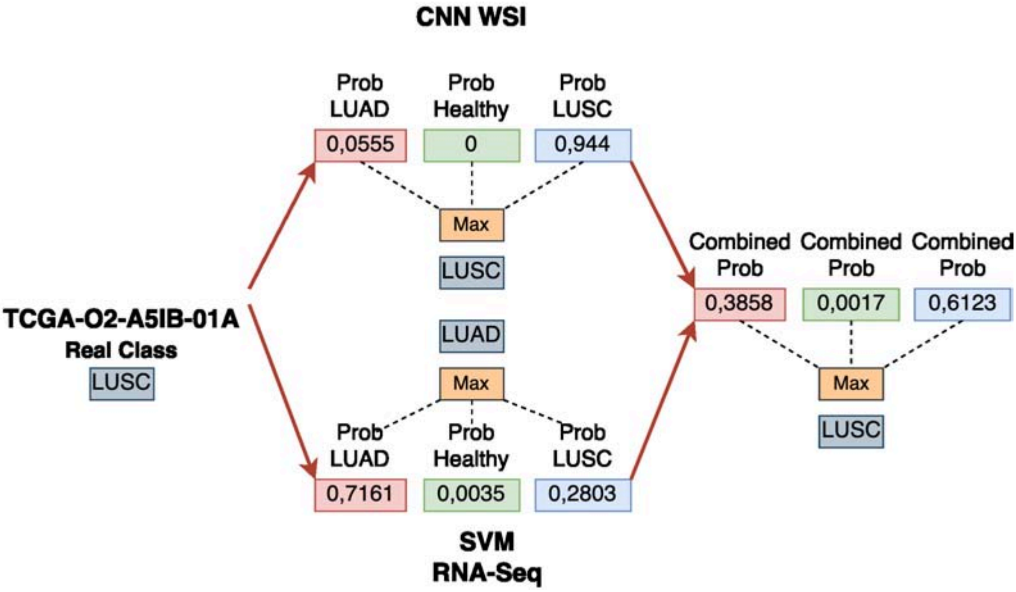


Figure 4.4: Example of the correct classification of a specific sample ID combining the probabilities. In the example shown, RNA-Seq classifier is providing a certain level of uncertainty between LUAD and LUSC classes, and due to the clear confidence of the CNN model for the LUSC class, the outcome of the fusion model provides the right diagnosis.

Table 4.3: Final Association Scores for the DEGs selected by mRMR. These scores have been obtained using the Open Targets platform [179]

mRMR Genes	Cancer	Lung Cancer	NSCLC	LUAD	LUSC
SLC2A1	0.53	0.27	0.30	0.08	-
NTRK2	1.0	0.30	0.30	0.06	0.04
TOX3	1.0	0.11	0.09	0.07	-
NXPH4	-	-	0.14	-	-
TFAP2A	1.0	0.68	0.70	0.67	-
KRT13	0.80	0.03	0.10	-	0.01

Table 4.4: Mean accuracy, F1-Score, AUC and standard deviation (in parenthesis) across the 10-Fold CV validation splits for each data type.

	F1-Score.(%)	**AUC**	**Acc.(%)**
WSI	83.39 (8.19)	0.947 (0.023)	86.03 (3.40)
RNA-Seq 3	90.57 (3.66)	0.978 (0.009)	90.67 (3.73)
RNA-Seq 6	93.67 (1.76)	0.987 (0.007)	93.70 (1.87)
RNA-Seq 10	94.05 (2.51)	0.990 (0.005)	94.12 (2.56)
Fusion 3	93.20 (3.18)	0.986 (0.005)	93.20 (3.17)
Fusion 6	95.19 (1.64)	0.991 (0.004)	95.18 (1.64)
Fusion 10	95.18 (1.61)	0.991 (0.005)	95.17 (1.62)

Table 4.5: Correct and erroneous predictions across the 950 samples when using 3, 6 and 10 genes

	WSI	**RNA-Seq 3**	**Fusion 3**
Correct	817	861	885
Misclassified	133	89	65
	WSI	**RNA-Seq 6**	**Fusion 6**
Correct	817	890	904
Misclassified	133	60	46
	WSI	**RNA-Seq 10**	**Fusion 10**
Correct	817	894	904
Misclassified	133	56	46

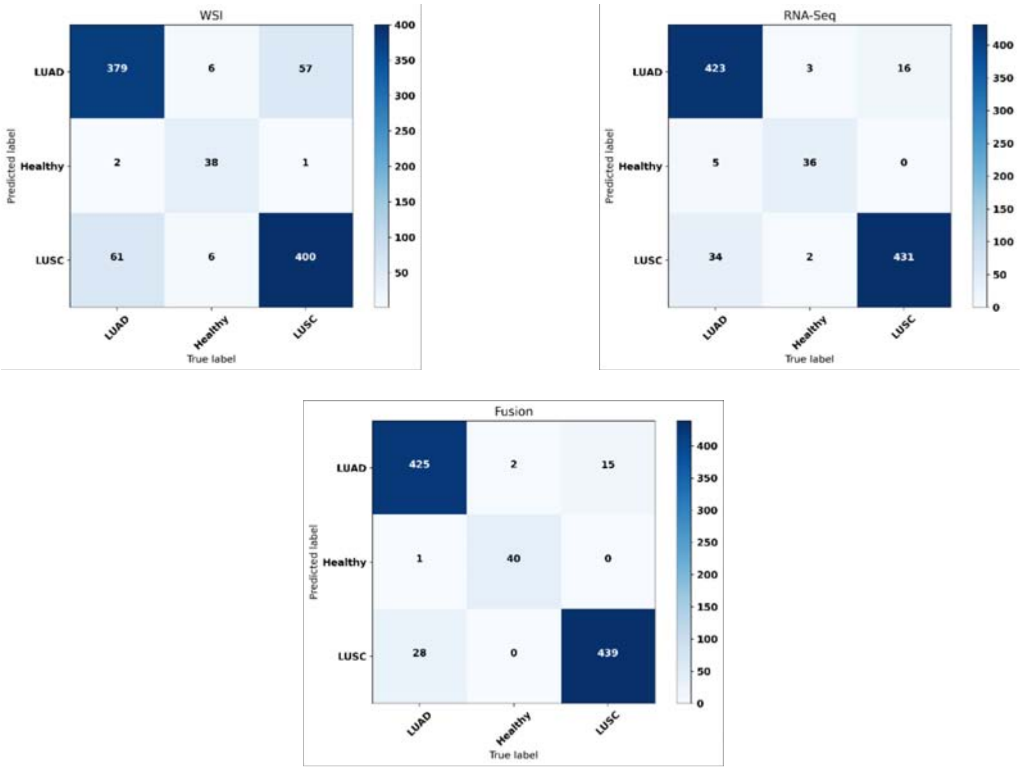
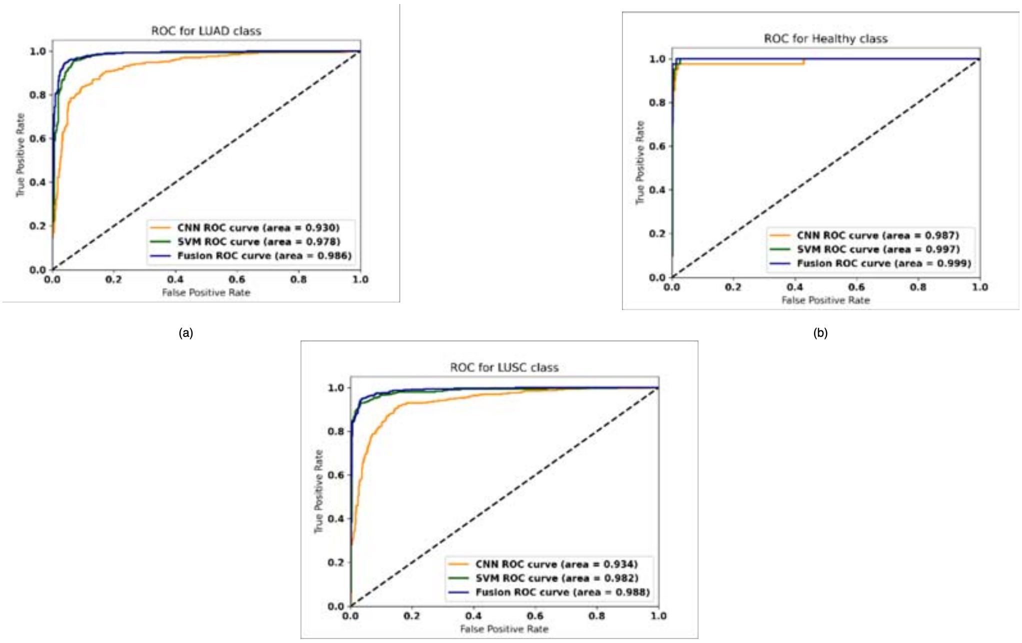


Figure 4.5: Confusion matrices obtained for the validation set in the 10-Fold CV by, (a) the CNN using WSI, (b) SVM using RNA-Seq data using 6 genes, (c) the fusion model using 6 genes. The accuracy and the f1-score is displayed under each confusion matrix.

Biological relevance of DEGs

The six DEGs found in our study are involved in different biological processes, usually associated with tumor development and cancer progression. Glucose metabolism can be found among these altered functions. Since cancer cells are highly proliferative, hypoxia often occurs when tumor cells



(a)

(b)

(c)

Figure 4.6: ROC curves obtained for the validation set in the 10-Fold CV for the CNN using WSI, SVM using RNA-Seq data using 6 genes, the fusion model using 6 genes for (a) LUAD, (b) Healthy and (c) LUSC classes. The Area Under the Curve for each classifier is displayed in the legend.

outstrip their vasculature. As a consequence, cancer cells show enhanced glycolysis as a means for energy production; this phenomenon is known as 'Warburg effect' [180, 181]. *SLC2A1*, also called *GLUT1*, codes for a glucose transporter that facilitates glucose transport across the plasma membrane [182]. *SLC2A1* overexpression was found in a series of solid tumors, including lung cancer [183]. More specifically, *SLC2A1* overexpression has been reported in bronchial brushing samples of NSCLC patients, and *SLC2A1* expression in NSCLC tissues was higher than in adjacent tissues [182]. In addition, *SLC2A1* overexpression has been linked to a poor prognosis in different cancer types, including NSCLC [184].

Some of the regulated genes encode for structural proteins, suggesting that cell architecture is also altered in NSCLC, which is related to loss of cell adhesion, invasiveness, migration, and metastasis. *KRT13* gene encodes for keratin (keratin-13), an intermediate filament protein expressed by epithelial cells in a cell-specific and differentiation-dependent manner [185]. It seems that enhanced *KRT13* expression in squamous-cell carcinomas could disturb the functions of cytoskeleton-cell junction desmosome and hemidesmosome protein complexes, consequently affecting cell adhesion and cell architecture, and indirectly affecting tumor behavior, neuroendocrine phenotypes, epithelial–mesenchymal transition (EMT) and stemness [186].

Modulation of genes that participate in signaling pathways such as *NTRK2* and *NXPH4* was also detected in our study. *NTRK2* is a member of the neurotrophic tyrosine kinase genes that encode for one of the Trk family proteins, TkrB. The NTRK family plays a role differentiation and maturation of the central and peripheral nervous system through activation of the PI3K-AKT and MAPK signaling pathways. However, *NTRK* gene fusions are found in solid tumors as oncogenic fusions responsible for growth and proliferation of cancer cells [187, 188]. Several studies have indicated that TkrB overexpression is oncogenic in several malignant tumors, including lung cancer. It might also be correlated with lymph node metastasis, vascular invasion and poor survival. Interestingly, Ozono et al. suggested that the binding of TkrB to one of its ligands, BDNF, promotes proliferating migratory and invasive phenotypes and cellular plasticity in LUSC [189], as previously reported for LUAD [190]. In contrast, although *NXPH4* up-regulation has been suggested in LUSC, its contribution to the disease remains unknown [191].

Furthermore, several of the DEGs were transcription factors, such as *TFAP2A and TOX3*. *TFAP2A* codes for the AP-2α transcription factor and many studies have been described it is markedly up-regulated, both in LUSC and LUAD tissues, compared with normal lung tissues. It seems correlated with poor prognosis, particularly among smokers [192], and it has been implicated in cancer proliferation, invasion, angiogenesis and EMT. Previous studies reported that *TFAP2A* promotes EMT by regulating TGF-β signaling in cancer cells and regulates tumor growth via hypoxia inducible factor-1a (HIF-1a) signaling in nasopharyngeal carcinoma and NSCLC [193]. For the *TOX3* gene, the protein produced contains an HGM-box, indicating that it may be involved in bending and unwinding of DNA and alteration of chromatin structure. Although its function remains unclear, it may be involved in various DNA-dependent processes [194]. TOX is a novel gene family that serves a pivotal function in human immunity. Recently, deregulated expression of TOX family members has been reported in a wide range of human cancer types. Notably, *TOX3* expression has been reported to be significantly increased in LUAD, compared with other pathological subtypes of lung cancer. Survival analysis demonstrated that elevated *TOX3* expression is significantly associated with improved progression-free and overall survival in patients with LUAD [195].

4.5 Conclusions

Promising results are obtained with each source of information, showing their potential to find cancer biomarkers. However, the proposed late fusion approach outperforms the results obtained by each classification model using RNA-Seq and WSIs in an isolated manner. It also reaches a more stable classification performance as observed in the experiments. This fusion model not only allows us to fuse the predictions from each classifier but also enables a prediction when some of the information is missing. The methodology used can also be universally applied to any kind of problem with heterogeneous data that presents missing information and the modularity of the system makes it easily scalable, so new classifiers for different types of data can be integrated with little effort.

The presented methodology represents an advancement in the creation of decision-making support systems that are applied to precision medicine, which can be used in a real-life scenario. With the integration of different sources of information, a more robust and complete prediction can be performed, similar to those situations in a hospital where different screenings are performed in order to diagnose a patient. Quick detection of any type of cancer in its early stage is crucial to improve the survival of the patient. Hence, accurate and fast methodologies, such as the one presented, can enhance the treatment of the patient.

In future work, we would like to test the proposed methodologies on other cancer types or diseases, in order to evaluate their general applicability. In addition, we would like to include more heterogeneous biological sources and domain knowledge, extending the flexibility of the model in face of real scenarios with different screenings performed, in the expectancy of an increase in the liability of the global diagnosis support system.

Non-small-cell lung cancer multi-omics and multi-scale classification

5

Source

Francisco Carrillo-Perez, Juan Carlos Morales, Daniel Castillo-Secilla, Olivier Gevaert, Ignacio Rojas, and Luis Javier Herrera. "Machine-Learning-Based Late Fusion on Multi-Omics and Multi-Scale Data for Non-Small-Cell Lung Cancer Diagnosis". In: *Journal of Personalized Medicine* 12.4 (2022), p. 601

Abstract

Machine learning techniques have provided a new framework for cancer diagnosis. By leveraging the information in the patient's data, a quicker and more accurate diagnosis can be provided. However, in most cases the cancer classification problem has been treated as a single-modality problem, not exploring the multi-scale and multi-omics nature of cancer data for the classification. In this work, we study the fusion of five multi-scale and multi-omics modalities (RNA-Seq, miRNA-Seq, Whole Slide Imaging, Copy Number Variation, and DNA-Methylation) by using a late fusion for non-small-cell lung cancer subtype prediction. We train independent classification models and explore the gains that are obtained by fusing their outputs incrementally, proposing a novel late fusion method based on stochastic gradient descent. The final classification model, using all modalities, obtains an F1-Score of 96.81 ± 1.07, an AUC of 0.993 ± 0.004 and an AUPRC of 0.980 ± 0.016, improving those results that each independent model obtains and those presented in the literature for this problem. In addition, the model is robust to missing modalities, making it more suitable to deploy on a real-world scenario. These obtained results show that leveraging the multi-scale and multi-omic nature of cancer data can enhance the performance of single-modality clinical decision support systems in personalized medicine, consequently improving the diagnosis of the patient.

5.1 Introduction

With the advances in computational methods, CDSS have been created for cancer detection using biological sources, achieving great results. Among the data sources used in literature we can find, for instance, WSI [131], gene expression data [112], CNV analysis [197], miRNA expression data [198] or DNA Methylation (metDNA) values [199]. By using these modalities independently, an accurate diagnosis can be performed. However, in their inner nature, they provide different biological information which may complement or regulate the information provided by the others. For instance, studies have shown that miRNAs regulate specific genes related to the proliferation of NSCLC [200, 201] or methylation and mutations patterns have been predicted using WSI [131, 202]. Therefore, exploring if the fusion of them can provide a more robust diagnosis using computational methods is of great interest for improving the prognosis of the patient.

In this work, we aimed to analyze the fusion of five heterogeneous modalities (WSIs, RNA-Seq, miRNA-Seq, CNV and metDNA) using a late fusion approach for the LUAD vs LUSC vs Control classification problem. We evaluated the improvements that can be obtained by fusing information and the modalities that are crucial to differentiate between the subtypes. In addition, a new late fusion optimization methodology is proposed for this problem, where the weights for the weighted sum of the probabilities are obtained by using a gradient descent approach that takes into account the performance of the fusion model in the classification.

5.2 Related work

Over the last few years, the potential of ML models using biological data for the diagnosis and prognosis of cancer patients has been shown. Specifically, all the aforementioned biological sources have been used for the creation of CDSS in lung cancer-related problems.

The use of gene expression data for lung cancer type classification has been explored in literature in recent years, especially for LUAD given that it is the most frequent NSCLC type. Smolander et al. reached a 95.97% of accuracy in the LUAD vs Control problem using coding RNA and employing a deep learning model [144]. Likewise, Fan et al. approached the same problem but using Support Vector Machines (SVM) with a 12 genes signature, obtaining an accuracy of 91% [145]. In addition, some works have been presented for the multiclass classification of lung cancer subtypes. Gonzales et al. presented a model for the classification of Small Cell Cung Cancer (SCLC), LUAD, LUSC, and Large Cell Lung Carcinoma (LCLC) by finding Differentially Expressed Genes (DEGs) and using them as input [147]. By employing RF as the feature selector and k-NN as the classification algorithm they obtained an accuracy value of 88.23%. Castillo-Secilla et. al. reached an accuracy of 95.7% using the Random Forest algorithm in the NSCLC subtypes classification task [203]. For the case of miRNA-Seq analysis, some works have been presented in the literature for lung cancer classification. Ye et al. presented a 10 miRNA signature for LUSC vs Control classification, reaching an F1-Score of 99.4% [198]. Yang et al. presented a miRNA signature for pathological grading in LUAD [204], reaching an accuracy of 66.19%. Also, miRNA has shown its potential for pan-cancer prognosis and treatment recommendation, including LUSC [119]. CNV data has also been used in literature for lung cancer classification. Qiu et al. presented a CNV signature for LUAD, LUSC and control classification formed by thirty-three genes reaching an accuracy of 84% in the validation set [197]. metDNA data has been used in literature for LUAD vs Control classification, reaching an accuracy of 95.57% by Shen et al. [205]. Also, the relation of DNA methylation-driven genes with LUSC and LUAD classes was studied by Gevaert et at. finding the clusters of methylation-driven genes that provided clinical implications [206]. Cai et al. test different feature selection algorithms in combination with different ML algorithms for the task of LUAD vs LUSC vs SCLC classification, reaching an accuracy of 86.54% on the task by using a panel of 16 CpGs sites [199].

Deep Learning (DL) has shown great potential for computer vision tasks, and therefore, its use combined with WSI has been explored in literature for NSCLC subtypes classification. Coudray et al. presented a Convolutional Neural Network (CNN) using tiles extracted from WSI for LUAD vs LUSC vs Control classification and mutation prediction, finally reaching an Area Under the Curve (AUC) score of 0.978 in the classification task [131]. By using manually labeled images by experts, Kanavati et al. presented a CNN model using transfer learning for the lung carcinoma vs control problem and obtained an AUC score of 0.988 [149]. Finally, other approaches have been presented where deep learning has been combined with more traditional statistics. Graham et al. used tiles extracted from the images and summary statistics to perform the classification between LUAD, control, and LUSC, reaching an accuracy value of 81% [150].

The fusion of the aforementioned sources has been explored in literature for various lung cancer problems, such as prognosis, grading prediction, or analyzing the relation between them. A Deep Neural Network (DNN) was developed by Lai et al. that combined gene expression and clinical data for prognosis prediction in NSCLC patients [140]. More novel techniques, such as autoencoders, have been explored in literature for the generation of a feature representation for a later fusion. Cheerla et al. used a deep learning-based model using miRNA, RNA-Seq, clinical, and WSI data for a pan-cancer prognosis prediction problem [115]. Similarly, Lee et al. used an autoencoder for the obtention of a feature representation using mRNA, miRNA, CNV, and metDNA for prognosis prediction [207]. For the problem of grading prediction, Long et al. proposed to use a late fusion methodology along with a gcForest model for predicting the stage of LUAD by fusion RNA-Seq, metDNA and CNV [174]. Authors reached an F1-Score of 88.9% on the task. Finally, in previous work, we showed that the fusion of WSI with RNA-Seq data improved the results obtained by each independent source for the LUAD vs LUSC vs Control problem [208].

As detailed, previous research has focused on the use of single modalities for the classification, obtaining great results with both molecular and imaging approaches. However, fewer works have been presented in literature performing a fusion of the information provided by these modalities, missing the opportunity to improve the classification performance and the knowledge acquisition from multiple biological sources. We propose to use the multi-modal information to enhance the classification performance for the sub-type identification, by leveraging the performance of independent classifiers and exploring the improvements that each source provides. A summary of the different works commented on for NSCLC classification problems is presented in Table 5.1.

Table 5.1: Summary of the works in literature for different NSCLC classification problems. SVM: Support Vector Machine; DNN: Deep Neural Network; RF: Random Forest; CNN: Convolutional Neural Network; k-NN: k-Nearest Neighbour; Acc.: Accuracy; AUC: Area Under the Curve

	Modalities	Problem	Model	Metrics	Results
Smolander et al. [144]	RNA-Seq	LUAD vs Control	DNN	Acc.	95.97%
Fan et al. [145]	RNA-Seq	LUAD vs Control	SVM	Acc.	91%
Gonzales et al. [147]	Microarray	SCLC vs LUAD vs LUSC vs LCLC	k-NN	Acc.	91%
Castillo-Secilla et al. [203]	RNA-Seq	LUAD vs Control vs LUSC	RF	Acc.	95.7%
Ye et al. [198]	miRNA-Seq	LUSC vs Control	SVM	F1-Score	99.4%
Qiu et al. [197]	CNV	LUAD vs Control vs LUSC	EN-PLS-NB	Acc.	84%
Shen et al. [205]	metDNA	LUAD vs Control	RF	Acc.	95.57%
Cai et al. [199]	metDNA	LUAD vs LUSC vs SCLC	Ensemble	Acc.	86.54%
Coudray et al. [131]	WSI	LUAD vs Control vs LUSC	CNN	AUC	0.978
Kanavati et al. [149]	WSI	Lung Carcinoma vs Control	CNN	AUC	0.988
Graham et al. [150]	WSI	LUAD vs Control vs LUSC	CNN	Acc.	81%

5.3 Material and Methods

Data acquisition and pre-processing

In this work, we have considered four molecular modalities and one imaging modality: RNA-Seq, WSIs, miRNA-Seq, Copy Number Variation and DNA Methylation Quantification. The data was collected from The Cancer Genome Atlas (TCGA) program [57], which is easily accessible from the GDC portal [154].

Biological and clinical information from 33 different cancer types is contained in TCGA and harmonization of all the samples has been performed by GDC. In most cases for each sample, various modalities are available (e.g. histology imaging, copy number variation, miRNA expression, gene expression, methylation beta values, etc.). Those Case IDs used in this work are available in a Github repository *. Table 5.2 shows the number of samples used per class and considered data modality.

Table 5.2: Number of samples per class for each data modality.

	WSI	RNA-Seq	miRNA	CNV	metDNA
LUAD	495	457	413	465	431
Control	419	44	71	919	71
LUSC	506	479	420	472	381
Total	1420	980	904	1856	883

To obtain unbiased results, a 10-Fold Cross Validation (10-Fold CV) training-test process was carried out in a stratified

* `https://github.com/pacocp/multiomic-fusion-NSCLC`

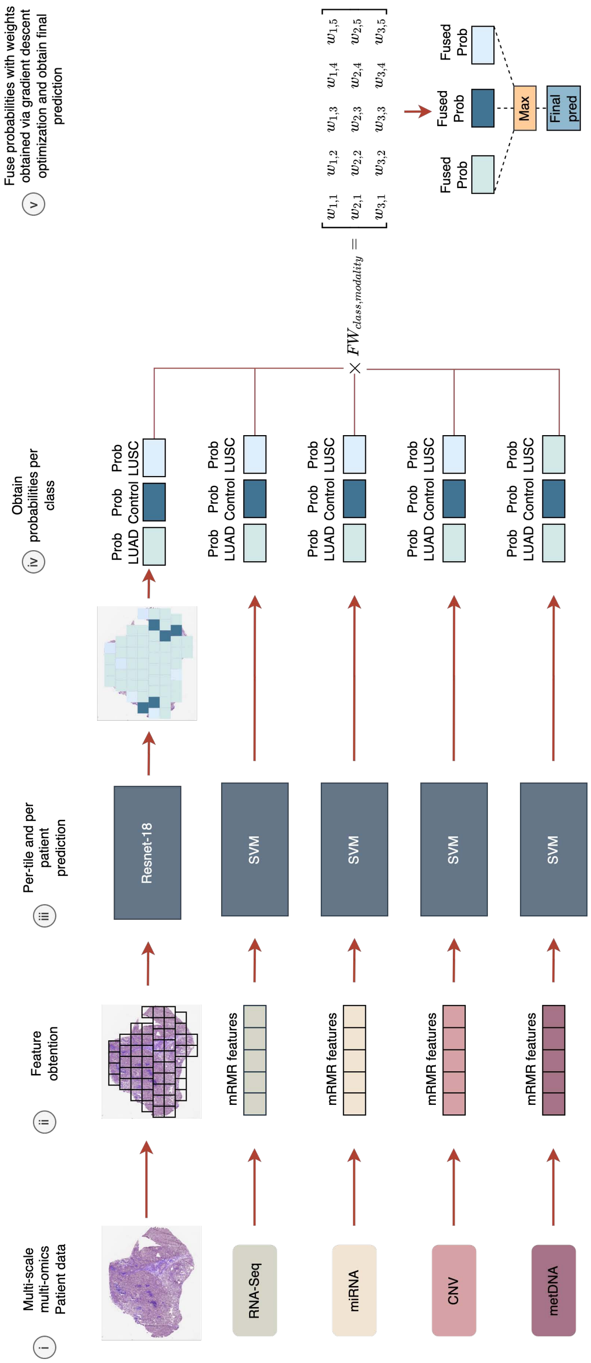

i
Multi-scale multi-omics Patient data
ii
Feature obtention
iii
Per-tile and per patient prediction
iv
Obtain probabilities per class
v
Fuse probabilities with weights obtained via gradient descent optimization and obtain final prediction
RNA-Seq
miRNA
CNV
metDNA
mRMR features
mRMR features
mRMR features
mRMR features
Resnet-18
SVM
SVM
SVM
SVM
Prob LUAD
Prob Control
Prob LUSC
$\times FW_{class,modality} = \begin{bmatrix} w_{1,1} & w_{1,2} & w_{1,3} & w_{1,4} & w_{1,5} \\ w_{2,1} & w_{2,2} & w_{2,3} & w_{2,4} & w_{2,5} \\ w_{3,1} & w_{3,2} & w_{3,3} & w_{3,4} & w_{3,5} \end{bmatrix}$
Fused Prob
Fused Prob
Fused Prob
Max
Final pred

Figure 5.1: Prediction pipeline for a given sample with multiple modalities. If missing information is present, the probabilities for that modality are zero. (i) Multi-scale and multi-omic data available for each sample is obtained. (ii) For the imaging modality, non-overlapping tissue tiles of 512x512 are obtained. For the molecular modalities, the features are obtained with the aforementioned preprocessing methodology (see Material and Methods section). (iii) Probabilities are computed for each modality and class. In the molecular modalities, the probabilities are returned by the machine learning model. For the imaging modality, the probabilities are obtained based on the number of tiles predicted per class divided by the total number of tiles. (iv) The late fusion model is applied using the previously obtained weights via gradient optimization, and the final prediction is obtained.

Table 5.3: Number of tiles obtained from the WSI per class.

	# Tiles
LUAD	100,841
Control	62,715
LUSC	92,584
Total	256,140

and patient-wise way over the whole dataset was performed. By doing it in a stratified way we are ensuring that we are maintaining the same proportion of classes across the splits, while in a patient-wise way we ensure that the samples from a given patient can only belong to one of the splits in each iteration, being it training or test. By doing so we are preventing any kind of information leakage between the splits. During each iteration, the training set was used for training the models, performing the biomarker identification, and for tuning the range of hyperparameters selected for the models, and once they were selected a final performance assessment was done on the test set. Different strategies were used for the hyperparameter selection depending on the data modality, which will be explained later in the manuscript.

WSI preprocessing

The python package openslide was used for the preprocessing of the obtained WSIs. We selected a magnification factor of 20x to obtain images with a sufficient resolution for the tile selection process (this magnification factor leaves images with a resolution of $\approx$ 10,000 x 10,000 pixels). For the tile selection process, we obtained 512x512 non-overlapping tiles of the whole image omitting those were there was a significant amount of background. To test this condition, we computed the mean value for the three color channels and if for the three channels the mean was greater than 220 we discarded that tile, as proposed by other authors in literature [131]. Otherwise, it was selected for further training. In Table 5.3 the final distribution of tiles per class can be observed.

Omic data preprocessing

To preprocess the RNA-Seq data, the KnowSeq R-Bioc package [203] was used to obtain the DEGs. The *DEGsExtraction* was used over $60,383$ genes from the training set in each split, similarly to other works that have been presented in literature [112, 161, 209]. As parameters, a Log_2 *Fold Chain* (LFC) value of 2, a p-value of 0.05, and a COV value of 2, were set. For the case of miRNA there is no need to apply a reduction of the number of features since TCGA provides information for 1881 miRNAs.

For metDNA and CNV values, the SciPy ecosystem's packages were used for the analysis and pre-processing [210]. TCGA contains information from 60.683 genes for CNV data, and $485,577$ known CpG sites for metDNA. Both sources contained missing values that were deleted, finally leaving with $46,585$ genes and $365,093$ CpG sites for the rest of pre-processing steps. In order to reduce the number of features, and to investigate the global difference in CNV and metDNA patterns among the three different groups (LUAD, LUSC, and Control), a two-tailed t-test was employed ($p \leq 0.001$), also using Bonferroni correction as a way to control for the family-wise error rate, as presented by Qui et al. [197]. Those genes and CpG sites that were significantly different in a number of the three two-tailed t-test comparisons (all of them for CNV and two out of the three for metDNA), and for which the difference of the mean was greater or equal to a given threshold (0.1 for CNV and 0.4 for metDNA), were selected in each split.

After performing the aforementioned pre-processing steps, the minimum Redundancy Maximum Relevance (mRMR) algorithm was used over the molecular data for obtaining the most important biomarkers in each modality, by obtaining the mRMR ranking [160]. Taking into account on every iteration we are using a different training split, which could lead to small variations in the biomarkers obtained each time.

Models selection and training

The Resnet-18 architecture was used for WSIs [166], using the pre-trained weights on Imagenet as the starting point [167] and normalizing the tiles using the mean and standard deviation from Imagenet. The last layer was adapted to the set of classes, and only this layer and the last residual block were trained (this last one was fine-tuned). For the selection of hyperparameters, a randomly selected 10% of each training set was used as validation in each split. The network was trained during 25 epochs using an early-stopping methodology where the accuracy in the validation set was monitored, saving the best weights for later use. Adam was used as the optimizer with the following hyperparameters: learning rate value of $1e^{-5}$, betas equal to $(0.9, 0.999)$ and epsilon equal to $1e^{-8}$, which were selected based on experimentation and results on the hyperparameter validation set. Once the per-tile model was obtained, for classifying a whole slide we followed a majority voting approach, similar to the one presented by Coudray et al. [131], where the final label was the most predicted class among all slide tiles. The training time of the model was around 14 hours using the ATCBIOSIUL cluster presented in Chapter 2.

For the rest of the molecular sources, different classification algorithms were tested, such as SVMS, k-Nearest Neighbors, or XGBoost. Finally, SVMs were chosen, since they obtained the best results in the training sets when performing the hyperparameter tuning and they have successfully used in literature for cancer classification with good results [112, 161, 162, 198, 204]. For tuning the SVMs hyperparameters a grid search CV was used over each training set. The only fixed parameter was the kernel, and we chose the Gaussian Radial Basis Function kernel based on the asymptotic behavior it has [172]. The search range of values for both C and γ was: $[2^{-7}, 2^{-5}, 2^{-2}, 2, 2^{4}, 2^{7}]$, and the features used were normalized between -1 and 1. The training time of the models was between 20-30 minutes using the ATCBIOSIMUL cluster presented in Chapter 2.

For implementing the classification models, the Python packages Pytorch [122] and Scikit-Learn [158] were used. In addition, the training of the Resnet-18 architecture was performed in an NVIDIA^{TM} RTX 2080 Super Graphics-Processing-Unit (GPU).

Probability fusion via weight-sum optimization

There exist two possibilities when applying a late fusion strategy: either to fuse the predictions [211] or the probabilities [177] returned by the classification models. The predictions can be fused by applying a voting scheme, where the most voted class among the different models is the one selected for the fusion model. On the other hand, with the probabilities, a more fine-grained fusion can be performed, since we have a probability percentage for each class. We chose this last option expecting a better performance based on previous results we have obtained on this problem when fusing RNA-Seq and WSI [208].

The approach for obtaining the probabilities differs between molecular and imaging modalities. For the molecular modalities, the probability for each class is obtained by using the methodology proposed by Wu et al. [173], based on a coupling method implemented in the SVM classifier that can be found in the Scikit-Learn python library [158]. For the imaging data, WSIs in this case, we need to manually compute them. Taking into account that we have the predictions for every tile in a given slide, the probabilities are computed as the number of tiles predicted for each class divided by the total number of tiles in the slide (see Eq. 5.1).

$$P^{CNN}(x, c_i) = \frac{\#TilesPredicted(x, c_i)}{\#SlideTiles(x)} \tag{5.1}$$

where x is the sample to be predicted and c_i is the given class: LUAD, Control, or LUSC.

Different approaches have been proposed in the literature to obtain the weights when there are different models and modalities. For instance, Dong et al. have proposed to compute the weights based on the performance of the classifiers without any normalization [174], while Meng et al. and Trong et al. have proposed to normalize the weights obtained based on the performance of the maximum accuracy or the maximum and the minimum accuracy respectively [175, 176]. Other approaches have consisted of simply multiplying the probabilities and the maximum was chosen for the prediction, by Depeursinge et al. [177]. In addition, we have previously proposed to compute the weights using stratified resampling sets using the performance of each model [208].

One drawback of the aforementioned approaches is that they are only taking into account the overall performance of the models. However, a classifier can be really good at discerning one or various classes but have low overall performance in comparison to the rest of the classifiers. In addition, the weights are computed only once and based on their individual performance, without taking into account how they performed in the classification task when they are fused. In this work, the probabilities from each model and class serve as input to an Artificial Neural Network (ANN), and the weights of the ANN are optimized using a stochastic gradient descent approach. By doing so, we can obtain a weight based on the performance of each classifier for each one of the classes, and where the weights change based on the performance of the fusion model in the classification task.

For the optimization of the weights, an ANN formed by a single linear layer was used in our study. The linear layer has 3x5 weights, which could be represented as the following matrix:

$$\begin{bmatrix} w_{1,1} & w_{1,2} & w_{1,3} & w_{1,4} & w_{1,5} \\ w_{2,1} & w_{2,2} & w_{2,3} & w_{2,4} & w_{2,5} \\ w_{3,1} & w_{3,2} & w_{3,3} & w_{3,4} & w_{3,5} \end{bmatrix} \tag{5.2}$$

where each row corresponds to a class (LUAD, Control, and LUSC respectively) and each column to a data modality (WSI, RNA-Seq, miRNA-Seq, CNV, and metDNA respectively). These weights are randomly initialized but fulfilling the condition that the row needs to sum up to one. After each backward pass, a softmax function is applied to the weights in order to maintain this condition.

Then, these weights are used to perform a weighted sum of the probabilities of each class (see Eq. 5.3), and the final predicted class is the one with the highest probability:

$$\begin{gathered} P^{c_i}_{Fusion} = P^{c_i}_{WSI} * w_{i,1} + P^{c_i}_{RNA} * w_{i,2} + \\ P^{c_i}_{miRNA} * w_{i,3} + P^{c_i}_{CNV} * w_{i,4} + P^{c_i}_{DNA} * w_{i,5} \end{gathered} \tag{5.3}$$

where c_i is the class (LUAD, Control or LUSC) and i is the index of the class in the weights matrix (see Eq. 5.2).

Once we have obtained the fused probabilities, the Cross-Entropy Loss is used as the loss function in order to optimize the weights for the classification task. By doing so, the optimization allows obtaining the combination of weights that maximizes the performance in the classification task. The Adam optimizer [212] is used for the optimization once the loss has been computed. A validation set of 10% was selected from each training set, in order to evaluate the performance of the fusion model during the optimization of the weights for 5 epochs.

This methodology allows us to easily deal with missing information, which is crucial when working with biological information given the high cost of performing all the screenings on a patient. If one of the data modalities is missing, its probability for each class will be zero and it would not affect the fusion (see Eq. 5.3). In Figure 5.1, an example of the prediction pipeline can be observed.

5.4 Results and Discussion

Performance of each data modality

For the late fusion strategy, we need to train independent models using each data modality. In the case of the molecular data, the number of features for each modality was selected based on having the lower number of features that provided the best performance for each independent model, by using the validation splits inside each training split in the 10-Fold CV process. Finally, 6 genes were selected for RNA-Seq, 9 miRNA for miRNA-Seq, 12 genes for CNV, and 6 CpGs sites for metDNA data.

The results that are obtained when using each source of information separately can be observed in Table 5.8, using all their available samples for the three classes classification problem (see Table 5.2). For the independent models, the higher results for the classification are obtained when using RNA-Seq and metDNA, followed by miRNA-Seq (see Table 5.8). These results are in accordance with previous studies in NSCLC. Qiu et. al. obtained an accuracy of 84% for CNV data [197]. Similarly, the results obtained by Cai et. al. (an accuracy of 86.54%) using metDNA are improved by those we have obtained [199]. For the case of WSI, the presented results are very similar to those obtained by Coudray et. al., an AUC of 0.978, and Graham et. al., an accuracy of 81% [131, 150]. For RNA-Seq, Castillo-Secilla et. al. reached an accuracy of 94.7% using SVMs, which is similar to our obtained performance [203].

Performance of late fusion with different number of sources

Once the models were trained, we tested the different improvements that can be obtained when adding new information, comparing the fusion of the sources in groups of two, three, four, and five. By doing so we are able to see how sources complement each other in terms of classification performance, and when it improves or worsens it. These results can be observed in Table 5.8. For the late fusion models, we used those samples that the data modalities in use have in common. The number of samples per class are provided in Tables 5.4, 5.5, and 5.6. The confusion matrices for the discussed fusion models are provided in Figure 5.2.

When fusing two sources, the highest performance in terms of classification metrics is obtained for the fusion of WSI-RNA-Seq, RNA-Seq-miRNA, and RNA-metDNA. Given that RNA-Seq, miRNA-Seq, and metDNA were the ones that achieved the highest performance independently, it is expected that their fusion provides an increase in the metrics. However, the fusion of WSI and RNA-Seq achieves great results in the classification, even though WSI is not among the sources with the best independent metrics. Therefore, WSI must be improving some of the RNA-Seq predictions, that might be on the wrong side of the prediction border of the probabilities.

Then, we moved to use three sources for the late fusion model. By adding miRNA data to the WSI-RNA-Seq fusion model the results obtained improved (from 94.69 ± 1.80 to 95.69 ± 1.76 in terms of F1-Score). The same happens when we include CNV or metDNA in the RNA-Seq-miRNA fusion model. RNA-Seq seems to be the most important source since it is included in those fusion models with high performance. In addition, the fusion of RNA-miRNA with other sources improves the classification over using RNA-Seq independently or with other sources.

Finally, we carried out experiments to observe if there is an improvement in the classification performance when using four or five sources. In this case, the only fusion that improves results over the fusion of three sources, in terms of the F1-Score and very similar results in the accuracy metric, is when we fuse all the biological sources. However, the improvement is really small and the standard deviation increases (95.69 ± 1.76 for WSI-RNA-Seq-miRNA and 95.82 ± 2.05 for the fusion of all

Table 5.4: Number of samples in common per class when we integrate two sources of information. WSI stands for Whole-Slide-Imaging, CNV stands for Copy Number Variation and metDNA for DNA Methylation.

Fusion	LUAD	Control	LUSC
WSI-RNA	442	41	467
WSI-miRNA	402	68	405
WSI-CNV	451	251	457
WSI-metDNA	415	71	356
RNA-miRNA	391	18	400
RNA-CNV	433	23	448
RNA-metDNA	398	14	348
miRNA-CNV	385	20	397
miRNA-metDNA	367	9	334
CNV-metDNA	441	52	361

Table 5.5: Number of samples in common per class when we integrate three sources of information. WSI stands for Whole-Slide-Imaging, CNV stands for Copy Number Variation and metDNA for DNA Methylation.

Fusion	LUAD	Control	LUSC
WSI-RNA-miRNA	381	15	389
WSI-RNA-CNV	419	23	437
WSI-RNA-metDNA	383	14	336
WSI-miRNA-CNV	376	20	383
WSI-miRNA-metDNA	356	9	319
WSI-CNV-metDNA	397	52	346
RNA-miRNA-CNV	369	5	377
RNA-miRNA-metDNA	346	4	314
RNA-CNV-metDNA	383	10	337
miRNA-CNV-metDNA	349	2	324

sources). For the rest of the fusion cases, the results obtained are similar to the highest reached when using three sources of information (see Table 5.8). Therefore, performing more screenings on the patient if you already have the biological sources that provided the best performance when using three sources is not necessary for an accurate diagnosis in this case.

Performance of the fusion models with missing information

Dealing with missing information is crucial when working with biological sources, given the high cost of some of the screenings. Therefore, we evaluated the effectiveness of the fusion model when some of the modalities are missing. In order to do so, for each fusion model the metrics were com-

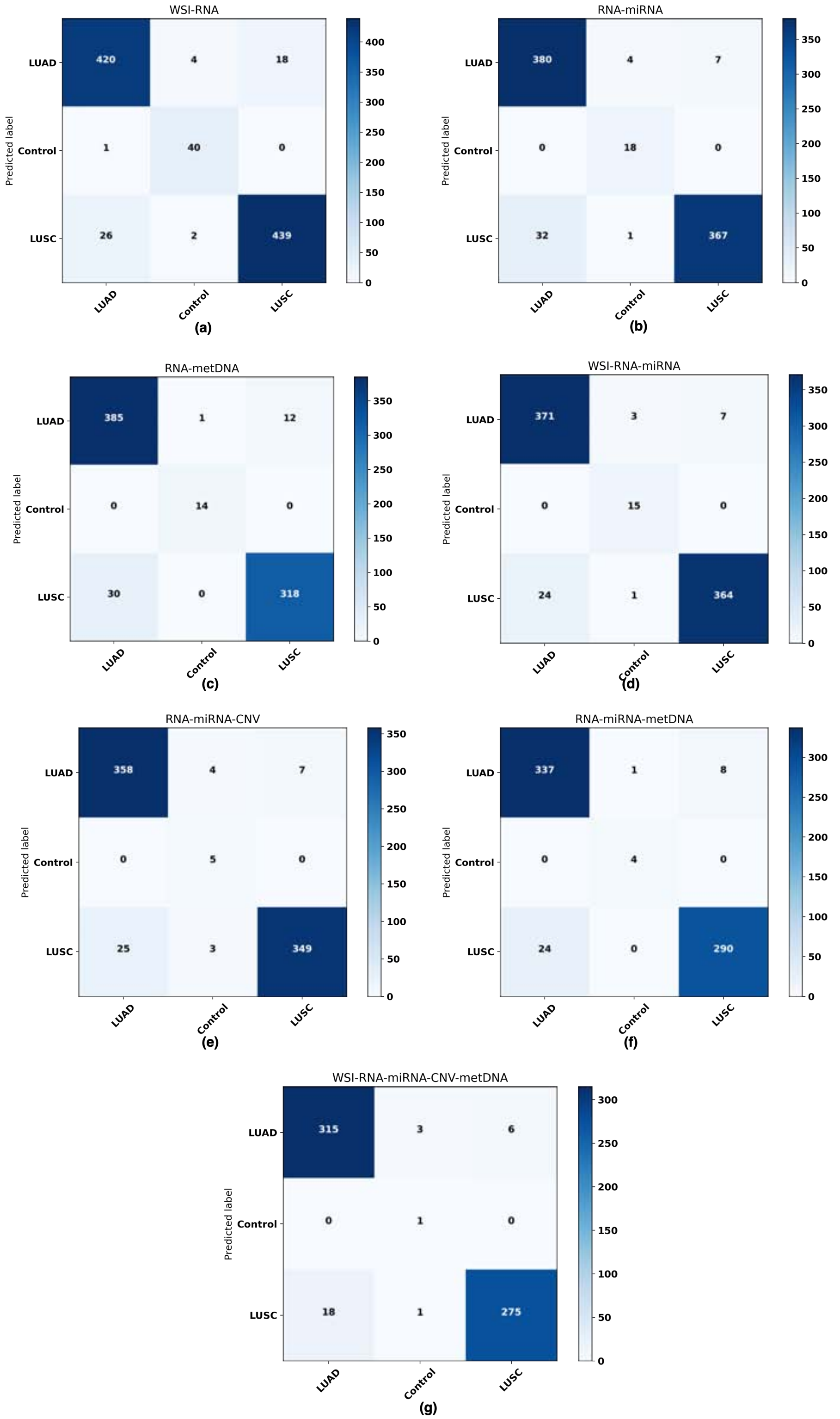


Figure 5.2: Confusion matrices obtained for different fusion models in the samples that the modalities have in common. WSI stands for Whole-Slide-Imaging, CNV stands for Copy Number Variation and DNA for DNA Methylation.

Table 5.6: Number of samples in common per class when we integrate four and five sources of information. WSI stands for Whole-Slide-Imaging, CNV stands for Copy Number Variation and metDNA for DNA Methylation.

Fusion	LUAD	Control	LUSC
WSI-RNA-miRNA-CNV	360	5	367
WSI-RNA-miRNA-metDNA	336	4	303
WSI-RNA-CNV-metDNA	369	10	326
RNA-miRNA-CNV-metDNA	333	1	304
WSI-RNA-miRNA-CNV-metDNA	324	1	294

Table 5.7: Ranges for the weights obtained for the five sources fusion and class in the 10 Fold-CV. The range presents the minimum and maximum values obtained with the optimization across the splits. WSI stands for Whole-Slide-Imaging, CNV stands for Copy Number Variation and metDNA for DNA Methylation.

Fusion		LUAD	Control	LUSC
WSI-RNA-miRNA-CNV-metDNA	WSI	[0.20, 0.33]	[0.21, 0.30]	[0.20, 0.28]
	RNA	[0.17, 0.30]	[0.11, 0.16]	[0.17, 0.22]
	miRNA	[0.15, 0.20]	[0.14, 0.20]	[0.17, 0.21]
	CNV	[0.16, 0.21]	[0.27, 0.34]	[0.17, 0.24]
	metDNA	[0.16, 0.21]	[0.11, 0.17]	[0.14, 0.22]

puted on all the samples available for the fused modalities, without restricting to those that the modalities have in common. In Figure 5.6, the F1-Score is presented for each fusion case predicting on all the samples that each modality has respectively (see Table 5.2). In addition, the confusion matrices for all the fusion models are presented in Figures 5.7, 5.8, 5.9, and 5.3.

Except for miRNA-Seq samples, the fusion that achieves the best performance is when fusing the five sources. However, the improvement is small in comparison with using four sources, so not having all of them does not excessively affect the classification performance. Fusing only CNV with metDNA, RNA-Seq or miRNA-Seq worsens the performance in comparison to the usage of them independently, which could be due to the imbalance of the classes (Table 5.2). The combination of metDNA and WSI also performs poorly, maybe due to the fact that it has been shown in the literature that WSI reflects information about the methylation patterns of human tumors [202], and therefore, they might not be complementing each other. However, in most cases including additional information improves the results that can be obtained by each independent source.

Table 5.8: Results obtained in the 10-Fold CV by every single modality and multi-modal fusion of the modalities in their common samples (see Tables 5.4, 5.5, and 5.6). For the case of four and five modalities fusion, AUC is omitted given the low number of control samples.

WSI	RNA-Seq	miRNA	CNV	metDNA	Acc. (std)	F1-Score (std)	AUC (std)	AUPRC (std)
X					88.56 (2.34)	88.57 (2.36)	0.965 (0.003)	0.940 (0.014)
	X				93.16 (1.87)	93.17 (1.82)	0.987 (0.007)	0.973 (0.028)
		X			92.31 (2.69)	92.34 (2.65)	0.976 (0.013)	0.961 (0.023)
			X		88.36 (1.34)	88.36 (1.34)	0.954 (0.009)	0.879 (0.025)
				X	93.21 (1.84)	93.19 (1.87)	0.972 (0.016)	0.957 (0.030)
X	X				94.65 (1.80)	94.69 (1.80)	0.991 (0.004)	0.979 (0.032)
X		X			92.59 (2.57)	92.60 (2.56)	0.987 (0.006)	0.982 (0.009)
X			X		90.26 (1.98)	90.20 (1.92)	0.974 (0.010)	0.962 (0.016)
X				X	92.79 (1.77)	92.80 (1.78)	0.983 (0.009)	0.979 (0.012)
	X	X			94.55 (1.83)	94.74 (1.70)	0.988 (0.007)	0.980 (0.017)
	X		X		91.81 (2.34)	92.12 (2.36)	0.978 (0.006)	0.953 (0.050)
	X			X	94.33 (1.81)	94.33 (1.79)	0.991 (0.007)	0.989 (0.009)
		X	X		91.00 (1.97)	91.36 (1.82)	0.973 (0.009)	0.944 (0.048)
		X		X	93.84 (2.88)	93.85 (2.88)	0.979 (0.015)	0.980 (0.015)
			X	X	90.15 (3.09)	90.28 (3.04)	0.968 (0.010)	0.947 (0.033)
X	X	X			95.55 (1.78)	95.69 (1.76)	0.985 (0.008)	0.990 (0.005)
X	X		X		93.99 (1.47)	94.00 (1.41)	0.982 (0.022)	0.974 (0.041)
X	X			X	94.70 (2.11)	94.73 (2.10)	0.987 (0.010)	0.990 (0.007)
X		X	X		93.84 (2.05)	93.97 (2.03)	0.974 (0.030)	0.977 (0.016)
X		X		X	94.23 (2.55)	94.23 (2.54)	0.975 (0.022)	0.986 (0.008)
X			X	X	93.50 (2.98)	93.52 (2.97)	0.981 (0.009)	0.978 (0.012)
	X	X	X		94.79 (1.76)	95.10 (1.72)	0.938 (0.059)	0.963 (0.050)
	X	X		X	95.05 (2.05)	95.10 (2.01)	0.967 (0.027)	0.989 (0.009)
	X		X	X	94.11 (1.76)	94.20 (1.74)	0.977 (0.012)	0.981 (0.010)
		X	X	X	94.11 (2.92)	94.36 (2.70)	0.975 (0.005)	0.966 (0.023)
X	X	X	X		95.22 (2.13)	95.47 (2.01)	-	0.987 (0.007)
X	X	X		X	95.53 (2.09)	95.62 (2.04)	-	0.989 (0.007)
X	X		X	X	95.22 (2.10)	95.30 (2.05)	-	0.986 (0.009)
X		X	X	X	94.71 (2.29)	94.9 (2.20)	-	0.978 (0.013)
	X	X	X	X	94.86 (2.19)	95.14 (2.06)	-	0.981 (0.010)
X	X	X	X	X	95.53 (2.20)	95.82 (2.05)	-	0.983 (0.012)

The final results obtained when fusing all the data sources is an F1-Score of 96.82 ± 1.07, an accuracy of 96.81 ± 1.07, an AUC of 0.993 ± 0.004, and an AUPRC of 0.980 ± 0.016. These results improved those aforementioned discussed and also reduced the standard deviation obtained across the splits. The receiver operating characteristic (ROC) curves (Figure 5.4 for the AUPRC metric and Figure 5.5 for the AUC metric) obtained show the performance of each individual modality and the fusion model over all the available samples for each one (all the samples in the case of the fusion model). The fusion model outperforms each modality for the three classes, showing the potential of using all the information. In addition, the fusion model reduces the number of misclassified samples for all sources, representing a reduction of the diagnosis error rate up to $\approx$ 8.6% in the best case and $\approx$ 1.6% in the worst case (see Table 5.9). The confusion matrix obtained over the whole dataset is presented in Figure 5.3 along with the weights obtained for each modality in the fusion (see Table 5.7).

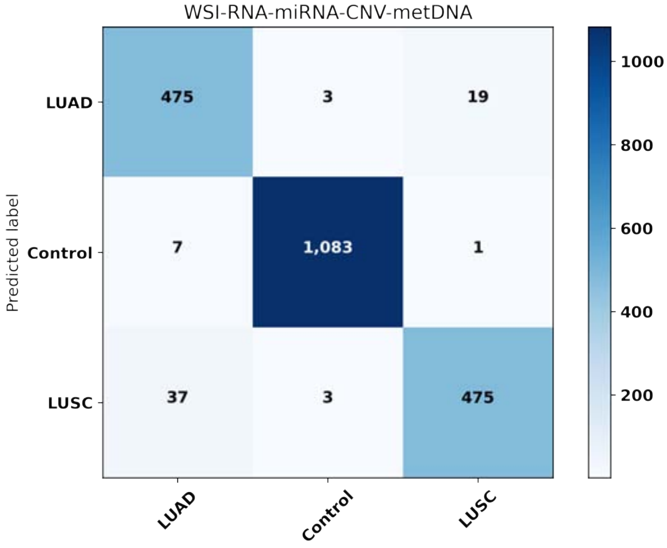


Figure 5.3: Confusion matrix for the fusion model using all modalities on all the available samples, without restricting to those that the modalities have in common. WSI stands for Whole-Slide-Imaging, CNV stands for Copy Number Variation and metDNA for DNA Methylation.

Comparison with previous work

The majority of the works presented in the literature for NSCLC subtypes and control classification have focused

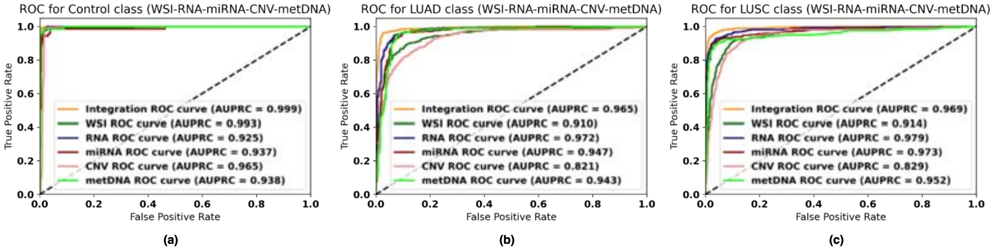


Figure 5.4: ROC curves for the fusion and individual models over all available samples for each modality showing the AUPRC metric. Figure 5.4 (a) ROC Curve for LUAD class. Figure 5.4 (b) ROC Curve for Control class. Figure 5.4 (c) ROC Curve for LUSC class.

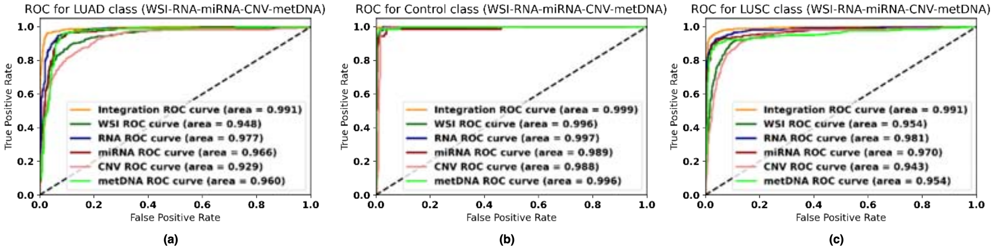


Figure 5.5: ROC curves for the fusion and individual models over all available samples for each modality showing the AUC metric. Figure 5.5 (a) ROC Curve for LUAD class. Figure 5.5 (b) ROC Curve for Control class. Figure 5.5 (c) ROC Curve for LUSC class.

on using a single data modality and mainly on a bi-class classification problem, and in some cases, a single train-test split was performed instead of a more robust k-Fold CV validation. Our fusion model outperforms or reaches the same results obtained by those works where a LUAD vs LUSC vs Control classification has been presented, and a summary is presented in Table 5.10. The fusion of information improves the results that Qiu et. al. obtained for CNV data (an accuracy of 84%) [197]. Similarly, the results obtained by Cai et. al. using metDNA are also improved (they obtained an accuracy of 86.54%) while reducing the number of CpG sites signature [199] and similar results are obtained compared to those presented by Castillo-Secilla et. al. using RNA-Seq (accuracy of 95.7%) [203]. For the case of WSI, the fusion also improved the results presented in the literature by Coudray et. al. [131] (an AUC of 0.978) reaching an AUC of 0.991. In the case of multi-omics fusion, we have not found works presenting methods for the NSCLC subtypes and control classification. However, in other NSCLC-related problems, the fusion of information has presented an enhancement in performance. Cheerla et. al. [115] showed that by fusing clinical, miRNA and

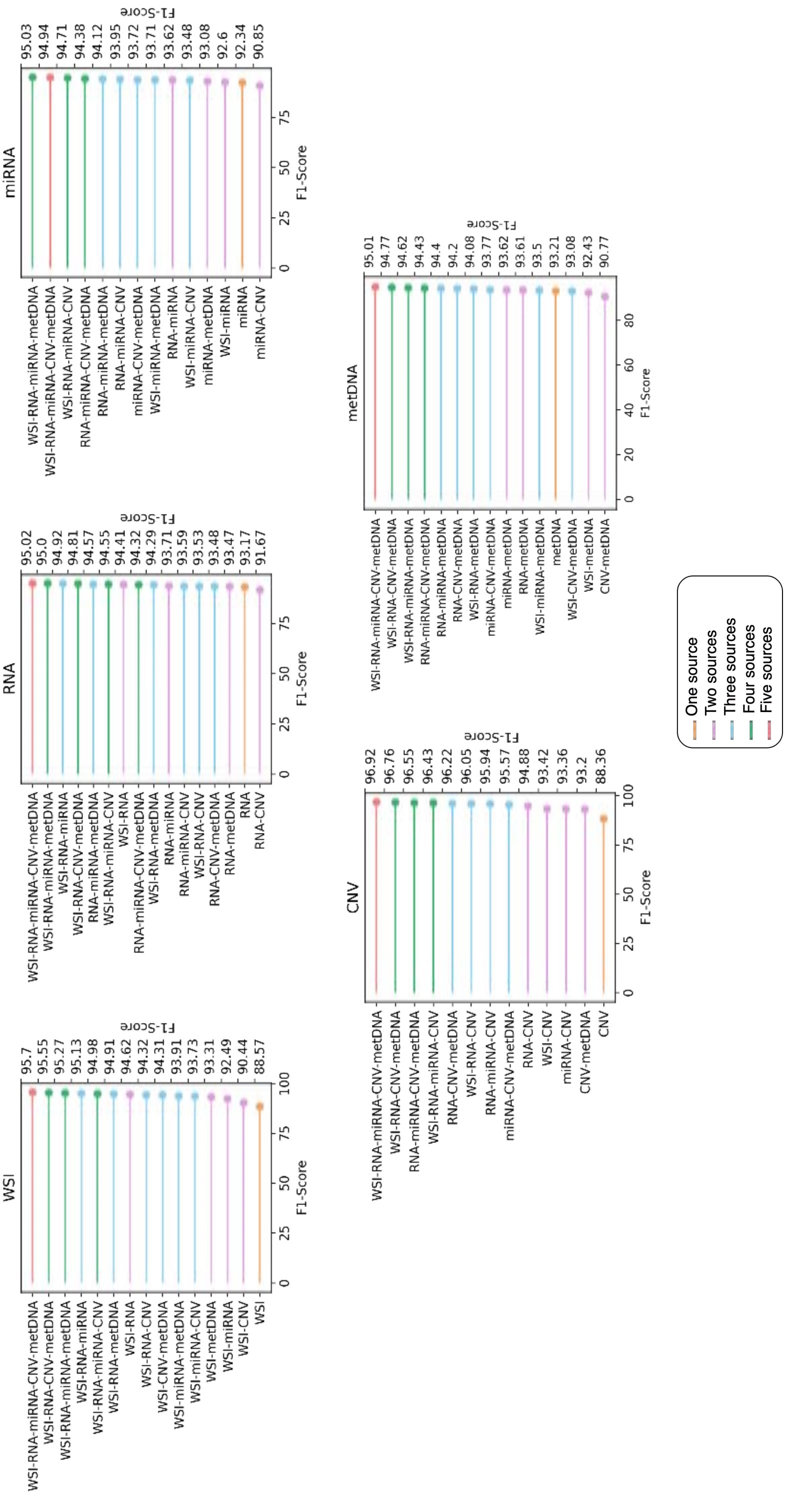


Figure 5.6: F1-Score obtained by each fusion model on the available samples for each modality, without restricting to those in common between the different modalities (see Table 5.2 check the number of samples per class). On the left Y-axis, the sources used in the integration are shown, while on the right Y-axis the F1-Score obtained by each integration can be observed. They are ordered from the highest F1-Score to the lowest. metDNA stands for DNA Methylation and CNV for Copy Number Variation.

Table 5.9: Correct and misclassified samples over the whole dataset for each data type and the fusion model using all modalities. RNA, CNV, and metDNA stand for RNA-Seq, Copy Number Variation, and DNA Methylation respectively.

	WSI	**RNA**	**miRNA**	**CNV**	**metDNA**
Correct	1232	913	834	1636	821
Misclassified	159	67	70	220	62
Fusion					
Correct	1328	929	857	1796	838
Misclassified	63	51	47	60	45
Absolute difference in misclassified error rate (#samples (%))	96 (6.5%)	16 (1.6%)	23 (2.6%)	160 (8.6%)	17 (2%)

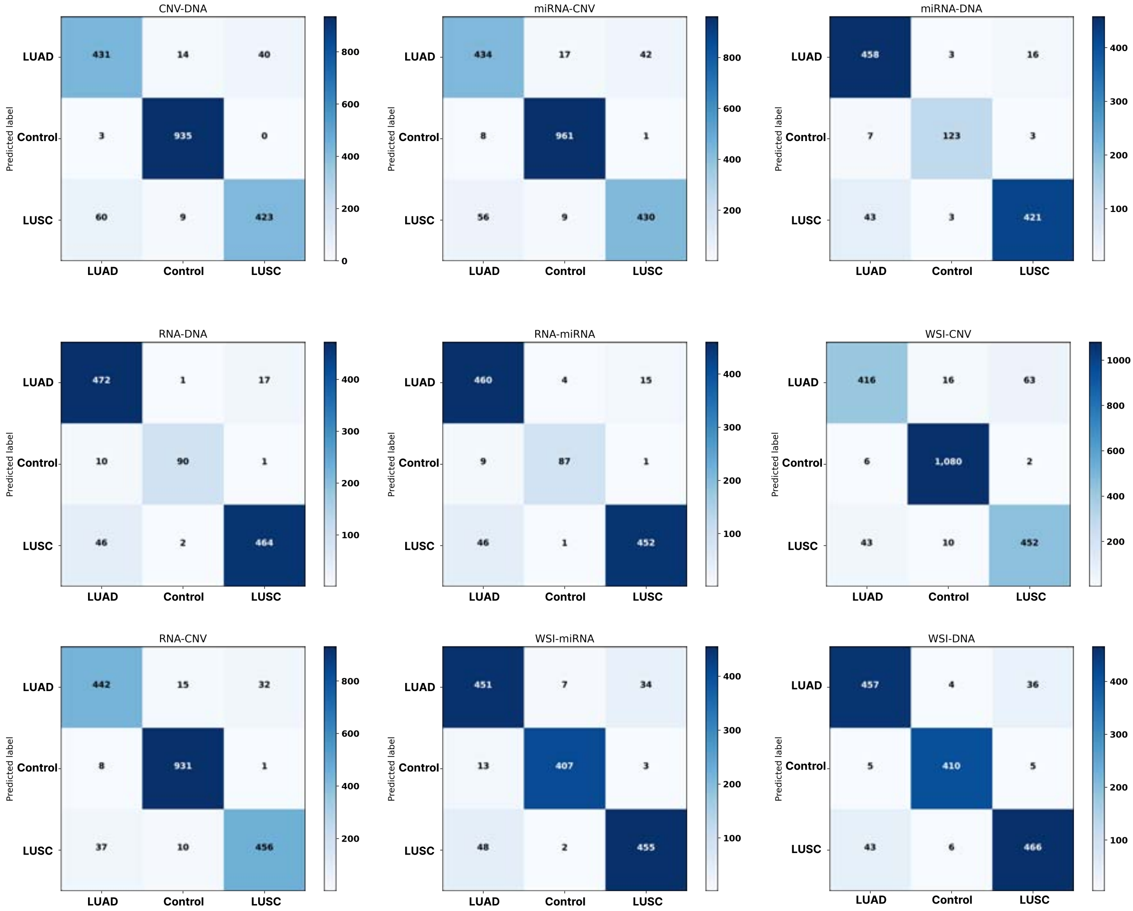


Figure 5.7: Confusion matrix for the two-sources fusion models using all modalities on all the available samples, without restricting to those that the modalities have in common. WSI stands for Whole-Slide-Imaging, CNV stands for Copy Number Variation and metDNA for DNA Methylation.

Table 5.10: Comparison of our fusion results with the available literature for LUAD vs Control vs LUSC. The results of the fusion model are on those available samples for the studied modality. Unfortunately, a direct comparison of fusion methods cannot be done given the lack of literature for this specific problem. The best results for each case are highlighted in bold.

	Modality	**Metric**	**Score**
Qui et al.[197]	CNV	Acc.	84%
Ours	CNV	Acc.	**96.93%**
Cai et al.[199]	metDNA	Acc.	86.54%
Ours	metDNA	Acc.	**95.01%**
Cai et al.[199]	metDNA	F1-Score	74.55%
Ours	metDNA	F1-Score	**95.01%**
Castillo-Secilla et al.[203]	RNA-Seq	Acc.	95.7%
Ours	RNA-Seq	Acc.	95%
Castillo-Secilla et al.[203]	RNA-Seq	F1-Score	**95.4%**
Ours	RNA-Seq	F1-Score	95.02%
Coudray et al.[131]	WSI	AUC	0.978
Ours	WSI	AUC	**0.991**
Graham et al.[150]	WSI	Acc.	81%
Ours	WSI	Acc.	**95.70%**

WSI data the performance was improved in LUAD prognosis prediction. Similarly, Lee et al. [207] improved the prognosis prediction by fusing the information of four sources (RNA-Seq, miRNA, CNV and metDNA) over each independent one. This same behavior is observed in our case for these sources.

When it comes to the relations between data modalities, our results highlight previously reported patterns. It has been presented in the literature that WSI can be used to predict mutation patterns or gene expression levels [131, 213], so the information provided may be completed with the one presented in RNA-Seq data. Similarly, when fusing three data modalities it was shown that including miRNA-Seq to the RNA-Seq-WSI fusion model improved the classification performance. miRNAs regulate specific genes related to the proliferation of NSCLC [200, 201], and therefore, might be complementing the information provided by RNA-Seq and WSI.

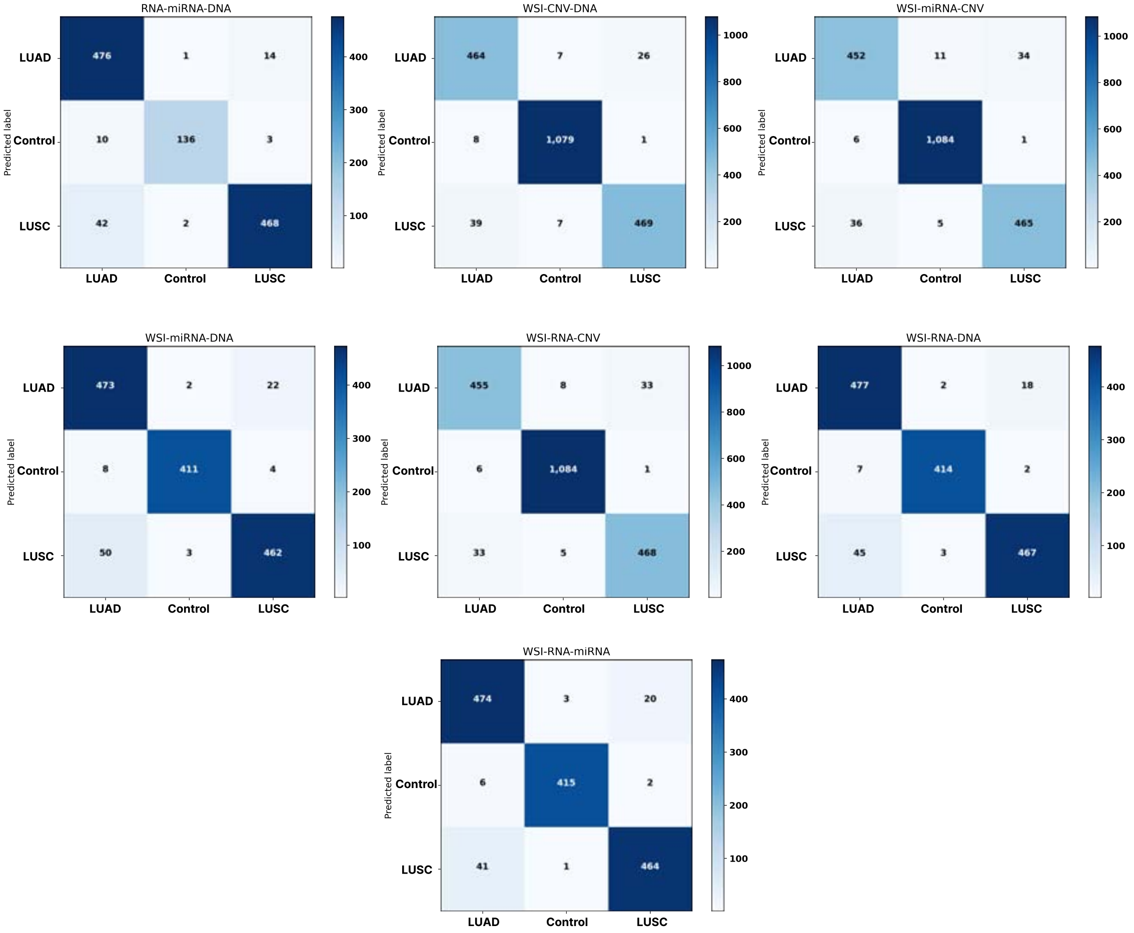


Figure 5.8: Confusion matrix for the three-sources fusion models using all modalities on all the available samples, without restricting to those that the modalities have in common. WSI stands for Whole-Slide-Imaging, CNV stands for Copy Number Variation and metDNA for DNA Methylation.

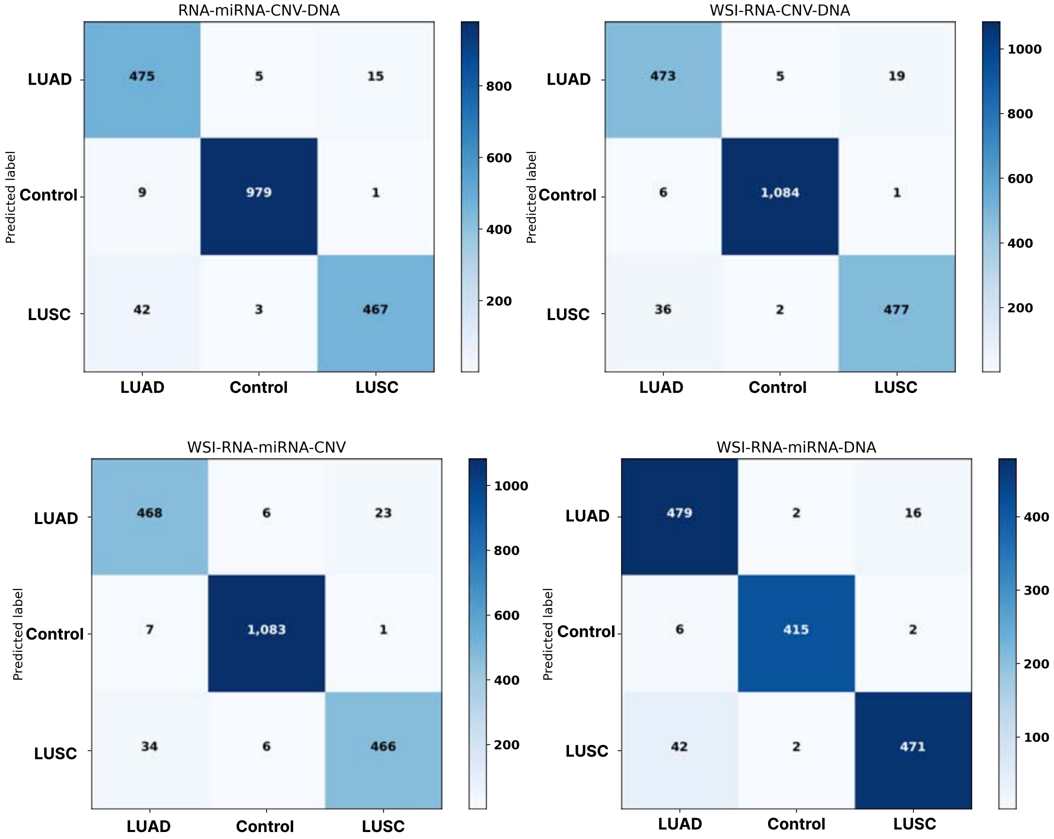


Figure 5.9: Confusion matrix for the four-sources fusion models using all modalities on all the available samples, without restricting to those that the modalities have in common. WSI stands for Whole-Slide-Imaging, CNV stands for Copy Number Variation and metDNA for DNA Methylation.

5.5 Conclusions

In this paper, we demonstrate the usefulness of fusing heterogeneous sources of biological information for NSCLC subtypes and control classification. In addition, we have proposed a new optimization methodology for weighting the classifiers in a late fusion strategy, effectively dealing with missing information and reaching good performance in the classification.

The fusion of the information outperforms the use of each independent source for the classification. Independently, RNA-Seq and metDNA achieve the highest performance in the classification. When performing the fusion, RNA-Seq is crucial for the classification problem and the addition of miRNA-Seq in combination with another data modality improves the obtained results. The best results are obtained when fusing the five sources of information reaching an F1-Score of 96.82 ± 1.07 when classifying all the available samples from all sources. However, there is not a huge increase in comparison with using three or four sources. The obtained results also highlight other reported patterns in the literature between data modalities that should be further studied. In addition, the methodology effectively deals with missing information, which is mandatory given that not all screenings are always performed on a patient. The presented methodology can be used in any diagnosing problem where heterogeneous sources of information are available, and it can be extended to any number of data sources. In future work, we would like to test the generalization capabilities of the proposed methodology for the classification of other cancer types or in other diagnosis-related problems and evaluate if the relations found between the different modalities apply to these other problems.

RNA-to-Image synthesis using deep generative models 6

Abstract

The study of how the different omics can affect a given tissue can unveil important mechanism in cancer biology, allowing a more in depth interpretation of tissue slides by pathologists or help to create better and more accurate CDSS. However, the number of samples available in multi-modal settings hugely limits this type of research. Generating synthetic samples of the different modalities arise as a solution to the scarcity problem, both to impute missing modalities or to increase the size of the dataset to create more powerful CDSS. Nevertheless, the majority of the approaches proposed in the literature focus on generating a single modality, not leveraging the information provided by others. Inspired by the recent advances in the text-to-image field, we proposed a solution for the RNA-to-image synthesis problem. We treat RNA-Seq data as the text and use it to generate corresponding WSI tiles. We proposed a novel approach using generative adversarial networks for generating healthy tissues. By doing so, we propose a solution to the scarcity problem, with a model that can be used to augment the data available in publicly available datasets, to impute modalities in non-multi-modal datasets available (where only the RNA-Seq data is present) or study the existing relation and effects of paired modalities on one another.

6.1 Introduction

Biomedical data has become increasingly multi-modal, which has allowed us to better capture the complexity of biological processes. In the multi-modal setting, several technologies can be used to obtain data from the same patient, providing a richer representation of their biological status and disease state. Currently, in clinical practice, demographic, clinical, molecular, and imaging data may be routinely collected on patients, making these data available for advancing the goals of precision medicine [7, 214]. For example, DNA and RNA-sequencing are now widely used for the characterization of cancer patients [215, 216]. Somatic mutation and gene expression profiles can be used to improve diagnosis, define disease subtypes, and determine the treatment regimen for cancer patients [131, 217]. Similarly, pathology WSI data are now more commonly available for secondary use. Specifically, digitization of hematoxylin & eosin (H&E) stained tissue sections from patients has become a key data source for training novel deep learning models. In the clinical setting, WSI is the cornerstone for a variety of tasks, such as primary diagnosis and treatment recommendations including the use of immunohistochemistry [133]. Specifically for oncology, morphological and texture changes can be observed in digitized tissue, reflecting the tumor microenvironment [133–135].

In particular, the relationship between genomic features and WSI image features has recently been demonstrated, with several studies showing that these two modalities are complementary. For example, genomic mutations, gene expression profiles, and methylation patterns have been predicted from WSI data using deep learning models [131, 202, 213]. Moreover, studies have shown that the integration of both modalities leads to an improvement in the performance of machine learning models for diagnostic and prognostic tasks in cancer [115, 131, 196, 218]. However, both modalities are not always available, due to financial or logistical constraints. Thus, opportunities for training models that require multi-modal data are missed, slowing down progress in advancing precision medicine [219, 220].

The scarcity of multi-modal data is a concerning problem in the machine learning community, especially in the context of recent successes for non-medical applications where huge amounts of data are available [77, 221]. To deal with this issue, a class of models in deep learning, generative models, have huge potential, by imputing data samples that are indistinguishable from real data, and as such creating synthetic data. Within generative models, GANs, VAEs, and more recently diffusion models have been widely used for multiple data generation tasks, obtaining exceptional performances in previous studies [222–224].

For non-medical applications, multi-modal data generation has gained interest in recent years thanks to the availability of large multi-modal data e.g. paired text and image data. Unsupervised learning methods such as GANs, transformers [225] and diffusion models [226] have been developed to leverage the relationship between these two modalities, enabling the generation of images based on their text description [224, 227–229], or showing image understanding and description capabilities [230]. The relation between these two modalities is similar to the relation between WSI images and genomic data since they are describing the same phenomenon from two different perspectives. As aforementioned, thanks to public multi-modal biomedical datasets such as The Cancer Genome Atlas (TCGA) [57] or the Genotype-Tissue Expression project (GTEx) [58], in-depth analyses can be carried out between these modalities, allowing to study diseases or biological processes based on their interactions.

However, these new methodologies rely on the use of deep neural networks for the analysis, which is widely known for being data hungry. Unfortunately, both modalities are not always available in public datasets. For example, the Genome Express Omnibus (GEO) database [231] has numerous RNA-Seq datasets available, but few datasets have the corresponding WSI images. Similarly, most medical centers have large archives of clinical slides, but not yet the means to generate matched gene expression data. New multi-modal datasets are being created to deal with these issues [232], yet the problem still occurs for most clinical data sets. While the multi-modal generation of data has been explored for natural images (text-to-image) [224, 227, 229, 233], the relation between WSI and gene expression needs yet to be explored for multi-modal synthetic data generation. Researchers usually focus on generating or imputing single modalities, without leveraging the information provided from other data types. However, by using both images and gene expression, the quality of the generated data can be significantly improved.

In this work, based on the success of text-to-image models in natural images, we questioned if we could provide a solution for the task of RNA-to-image synthesis. First, we are going to generate healthy tissue tiles from two different tissues, lung, and brain cortex, using traditional GANs (see Figure 6.1 B). Then, we present a novel approach for RNA-to-Image synthesis by infusing a GAN architecture with the gene expression profile of the patient to generate synthetic tiles (see Figure 6.1 C).

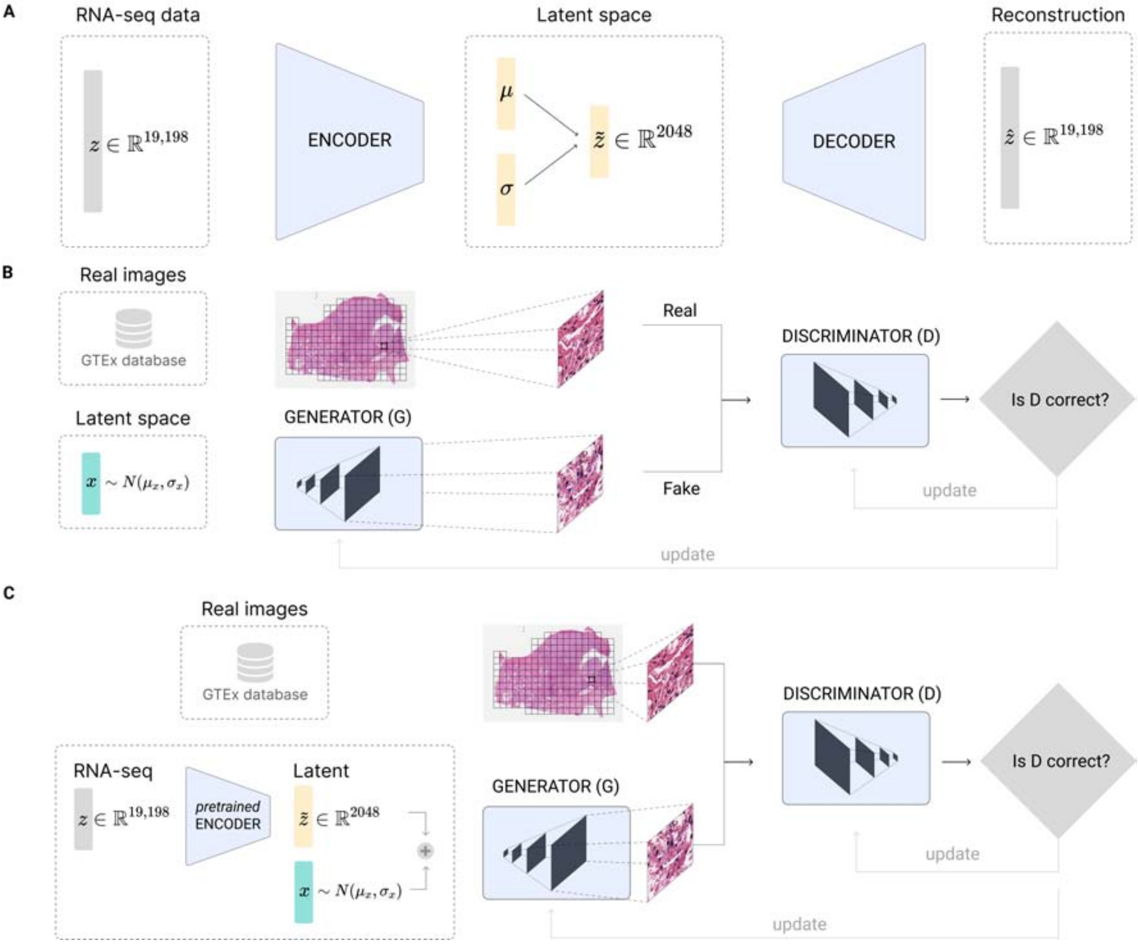


Figure 6.1: Model architecture for gene expression, WSI, and combined data using VAE and GANs. Panel A: βVAE architecture for the generation of synthetic gene expression data. The model uses as input the expression of $19,198$ genes. Both the encoder and the decoder are formed by two linear layers of $6,000$ and $4,096$ respectively. The latent μ and σ vectors have a feature size of $2,048$. Panel B: GAN architecture for generating tiles by sampling from a random normal distribution. The architecture chosen was a Deep Convolutional GAN (DCGAN) [101], using as input a feature vector of size $2,048$. The final size of the tiles generated is 256×256, the same as the size of the real tiles. Panel C: RNA-GAN architecture where the latent representation of the gene expression is used for generating tiles. The gene expression profile of the patient is used in the βVAE architecture to obtain the latent representation. Then a feature vector is sampled from a squeezed random normal distribution (values ranging between $[-0.3, 0.3]$) and added to the latent representation. A DCGAN is trained to use this vector as input and generate a 256×256. The discriminator receives synthetic and real samples of that size.

6.2 Related work

While the RNA-to-image synthesis task has not been approached in the literature, several studies have focused on the generation of single-modality synthetic data for both RNA gene expression and WSI data. For example, the generation of gene expression data has been mainly used in the context of data imputation and has been researched by leveraging the latent space of VAEs. Qiu et al. showed that $beta$VAEs, a special case of VAEs, can impute RNA-Seq data [234]. Similarly, Way et al. proposed a VAE trained on pancancer TCGA data, that is able to encode tissue characteristics in the latent space and also leverages biological signals [235]. Also, Vinas et al. presented a model that could generate synthetic gene expression profiles that closely resemble real profiles and capture biological information [236]. The generation of high-quality WSI tiles has also been researched in recent years given the success of GANs in generating natural images [102, 103]. For example, Quiros et al. showed that GANs are able to capture morphological characteristics of cancer tissues, placing similar tissue tiles closer in the latent space, while generating high-quality tiles [105, 237].

6.3 Material and Methods

Data

The Genotype-Tissue Expression (GTEx) project was used to obtained the pared WSI and RNA-Seq. We collected the RNA-Seq and WSIs from brain cortex, lung, pancreas, stomach and liver tissues from the GTex database. There were a total of 246 samples of brain cortex tissue, 562 samples of lung tissue, 328 samples of pancreas tissue, 356 of stomach tissue, and 226 samples of liver tissue. To validate the generalization capabilities in generating tiles from the gene expression of other cohorts, the GEO series 120795 was used [238].

RNA-Seq data preprocessing

Gene expression data from the GTEx project contains a total of $56,201$ genes. This number would require huge computational capabilities, and it difficults the training of the machine learning models. Therefore, we reduced the feature dimension and obtained the expression of $19,198$ protein-coding genes for further experiments. The data was log-transformed, and the z-score normalization was applied to the gene expression using the training set data, in order to not include the validation or the test set information on the normalization process. In the generalization experiments, the gene expression from lung and brain cortex tissue of the GEO series 120795 was used. However, not all the previously selected protein-coding genes were among those sequenced in this dataset. Therefore, for those missing in this external cohort, we initialized them as zero for the generation of the tiles. Data were normalized using the mean and standard deviation from the training set of the GTEx data and log transformed.

WSI data preprocessing

WSIs were acquired in SVS format and downsampled to 20× magnification ($0.5\mu m$ px^{-1}). The size of WSIs is usually over $10k \times 10k$ pixels, and therefore, they cannot be directly used to train machine learning models to generate synthetic data. Instead, tiles of a certain dimension were taken from the tissue, and these are used to train the models, which is consistent with related work in state-of-the-art WSI processing [131, 239, 240]. In our work, we took non-overlapping tiles of 256×256 pixels. Firstly, a mask of the tissue in the higher resolution of the SVS file was obtained using the Otsu threshold method [241]. Tiles containing more than 60% of the background and with low contrast were discarded. A maximum of $4,000$ tiles were taken from each slide. For the preprocessing of the images we relied on the python package openslide [155], which allows us to efficiently work with WSI images. The tiles were saved in an LMDB database using as an index the number of the tile. This approach enables us to reduce the number of generated files, and structure of the tiles in an organized way for a faster reading while training. Tiles containing pen marks or other artifacts were filtered during the reading phase.

beta-VAE for encoding RNA-Seq and generate synthetic samples

To reduce the dimensionality of the RNA-Seq data, we decided to use a β-VAE architecture. For the RNA-to-image synthesis of healthy samples, we empirically determined to use two hidden layers of $6,000$ and $4,096$ neurons each for both the encoder and the decoder, and a size of $2,048$ for the latent dimension. Given that we were going to use the latent representation for the generation of the tiles, we followed the same dimensionality as the output of the convolutional layers of state-of-the-art convolutional neural networks [166]. We used batch norm between the layers and the LeakyReLU as the activation function. A $\beta = 0.005$ was used in the loss function. We used the Adam optimizer for the training with a learning rate equal to 5×10^{-5}, along with a warm-up and a cosine learning rate scheduler. We trained the model for 250 epochs with early stopping based on the validation set loss, and a batch size of 128. A schema of the architecture is presented in Figure 6.1 A. We divided the dataset in 60-20-20 % training, validation and test stratified splits. We trained two different models, one for brain cortex and lung tissue data, and the other with all the tissues described in previous subsections (lung, brain cortex, stomach, pancreas, and liver).

GAN and RNA-GAN for the generation of healthy synthetic samples

For the generation of synthetic samples, we propose two architectures in order to compare the advantage of using the RNA-Seq profile of the patient for the generation of the tiles.

Firstly, we used a normal GAN (described in Chapter 2). Specifically, we trained two Deep Convolutional GANs [101], one per tissue, by sampling from a normal random distribution (scheme depicted in Figure 6.1 B). We sample a different number of tiles per image for the training of the network, finally selecting 600 tiles per image because the quality of the image and the artifacts were highly improved by augmenting the number of tiles. We used the Adam optimizer for both the generator and the discriminator, with a learning rate equal to 1×10^{-3} for the generator, a learning rate equal to 4×10^{-3} for the discriminator, and betas values $(0.5, 0.999)$ in both cases. Data augmentation such as random vertical and horizontal flips were used during training. The brain tissue GAN was trained during 39 epochs while the lung tissue GAN was trained during 91 epochs. For the training of the GANs, the Python package Torchgan was used [242].

Secondly, we present a novel architecture, called RNA-GAN for RNA-to-image synthesis (see Figure 6.1 C). We combined the pretrained βVAE with the DCGAN architecture, using the encoding in the latent space as the input for training the generator. To generate different tiles from the same gene expression profile, we sample a noise vector from a narrowed random normal distribution (values ranging between $[-0.3, 0.3]$) and add it to the latent encoding. Therefore, the input to the generator would be:

$$\tilde{x} = q_\theta(z|x) + N(0, 1) \tag{6.1}$$

We trained two DCGANs, one per tissue, and the pipeline is depicted in Figure 6.1 C). We finally selected 600 tiles per image to train the generator. We used the Adam optimizer for both the generator and the discriminator, with a learning rate equal to 1×10^{-3} for the generator, a learning rate equal to 4×10^{-3} for the discriminator and betas values $(0.5, 0.999)$ in both cases. Data augmentations such as random vertical and horizontal flips were used during training. The brain tissue GAN was trained during 24 epochs while the lung tissue GAN was trained during 11 epochs. For the training of the GANs, the Python package Torchgan was used [242].

To validate the generalization capabilities of the trained model, the GEO series 120795 was used. It contains gene expression profiles from healthy tissues, where we took the expression of lung and brain cortex tissues. For obtaining machine learning performance metrics, one hundred images were generated per tissue and obtain from real data. Then, a Resnet-18 was trained in the real data from scratch using 10 epochs and early stopping based on a 20% of data as validation set. A learning rate value of $3e^{-5}$ and AdamW optimizer were used. Finally, the model was tested on the GEO synthetically generated data, and accuracy, F1-Score and AUC was computed.

To evaluate the quality of the synthetic tiles, we presented a form to expert pathologists. The pathologists were not informed that some presented tiles were synthetic, to omit any kind of biases in the evaluation. Instead, we informed the pathologists that these tiles were going to be used to create machine learning classifiers, and we wanted to evaluate their quality for this task. Three questions were asked to the experts:

1. Is the tile from brain cortex or lung tissue?
2. Quality of the morphological structures: Being 1 very bad and 5 very good, how would you rate the morphological features present in the tile for an assessment of the tissue?
3. Do you find artifacts in the image? (e.g. image aberrations) (Yes/No)

The training time of the GAN model was around 12 days for the lung tissue and 8 days for the brain tissue using the Sherlock cluster presented in Chapter 2. The training time of the RNA-GAN model take around 48 hours for the brain tissue and 58 hours for the lung tissue using the same cluster. The models were trained using mainly NVIDIA V100 GPUs.

6.4 Results and Discussion

A β-VAE model discriminates between tissues and can generate synthetic multi-tissue expression profiles

As a first step, we aimed to create an accurate, distinguishable latent representation of healthy multi-tissue gene expression using a β-VAE architecture (Figure 6.1 A). The β-VAE model was able to accurately reconstruct the gene expression by forwarding the latent representation through the decoder and obtaining a mean absolute error percentage of 39% (RMSE of 0.631) on the test set for multiple tissues (Figure 6.2 C). To verify that the latent representation learnt by the βVAE accurately maps to the different tissues, the UMAP algorithm [243] was used to visualize the real gene expression data as well as reconstructions of latent representations on the test set. For lung and brain samples, two separated clusters can be distinguished, showing how the model is characterizing the two tissues in the latent space (Figure 6.2 A, "Real" versus "Reconstruction").

To further validate the learned latent space, we tested what happens when interpolating data in it. By interpolating in the latent space, we should be able to "transform" a randomly drawn sample to a gene expression profile that looks like it originated from one of the tissues (i.e. synthetic gene expression generation). To do so, we need to calculate the cluster centroid vector over the real data latent representations of the desired tissue and add this centroid vector to randomly drawn samples from the βVAE latent distribution. This procedure allows us to generate synthetic gene expression data that look like real brain or lung gene expression data. When projecting these synthetic samples in the UMAP space, they indeed fall in the same clusters as the original data (Figure 6.2 A, "Generated" versus "Real").

We can also perform other operations in the latent space. For example, we should be able to "shift" the gene expression from one tissue into what it would look like if it originated from another tissue. In this case, we need to add the difference vectors between the cluster centroids of the respective tissues to the latent representation of a given sample gene expression. For example, we can shift a real brain gene expression profile to a lung gene expression profile and vice versa. Visualizing these new samples in the UMAP space verifies that these operations can indeed be successfully performed (Figure 6.2 B). Next, the representation capabilities of the βVAE can also be extended to multiple tissues, showing a diverse representation with well-differentiated clusters, and maintaining the generative capabilities across the multiple tissues (Figure 6.2 C).

GANs generate quality synthetic WSI tiles preserving real data distribution differences

Next, we developed a traditional GAN model to generate synthetic WSI tiles for brain cortex and lung tissue. The model was able to generate good quality images, preserving the morphological structures, and showing little artifacts (Figure 6.3 A). In some tiles, checkerboard artifacts are noticeable, which is a known problem in GANs [244].

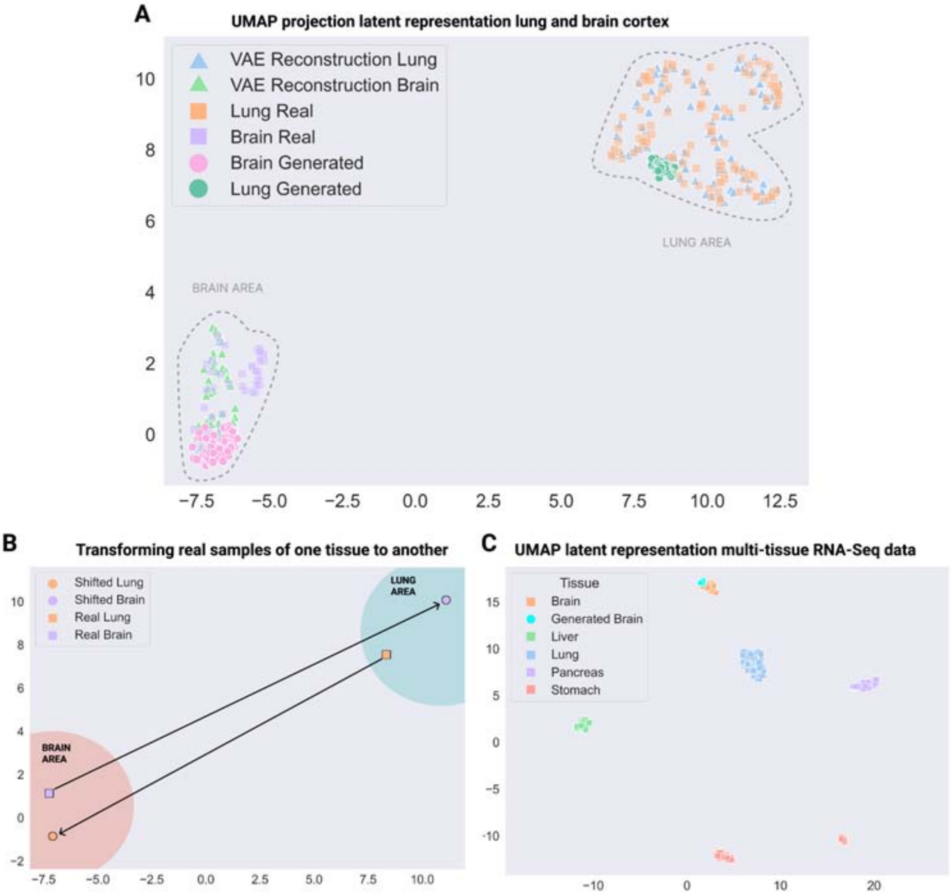


Figure 6.2: UMAP visualization of β-VAE embedding of multi-tissue expression profiles. Panel A: UMAP visualization of the real and reconstructed gene expression profiles of lung and brain cortex healthy tissue. Generated gene expression profiles, by sampling from the latent space and interpolating to the respective tissue, is also plotted showing the generative capabilities of the model. Panel B: Shifting real gene expression profiles between the two tissues. The latent representation of all the available samples is obtained, and the difference vectors between the clusters centroids are computed. Panel C: UMAP visualization of real gene expression profiles of multiple tissues and generated one from brain cortex tissue.

Despite the artifacts, the main cell types can be observed in the tiles, such as epithelial, connective, and muscle tissue. In addition, there is a clear distinction between the tiles generated for the brain cortex and the lung, preserving the characteristics of the corresponding real tiles. Specifically, the brain cortex tissue is grouped in a set of layers that form a homogeneous and continuous layer (the outer plexiform layer, outer granular layer, outer pyramidal cell, inner granular layer, inner pyramidal layer, and polymorphous layer) [245]. These characteristics can be observed in the synthetic brain tiles, i.e. they appear more homogeneous and contain less white spaces in comparison to the synthetic lung tissue tiles. The synthetic lung tissue tiles also present the characteristics of real tiles, showing the terminal bronchioles, respiratory bronchioles, alveolar ducts, and alveolar sacs in some cases.

To test if the generated tiles have the same distribution as the real ones, the feature vector outputted from one of the last convolutional layer of an Inception V3 network pretrained on Imagenet was obtained for the 600 generated tiles. Then, these feature vectors were projected and visualized using the UMAP algorithm, showing a similar distribution between the tissues for both real (Figure 6.3 B) and synthetic samples (Figure 6.3 C).

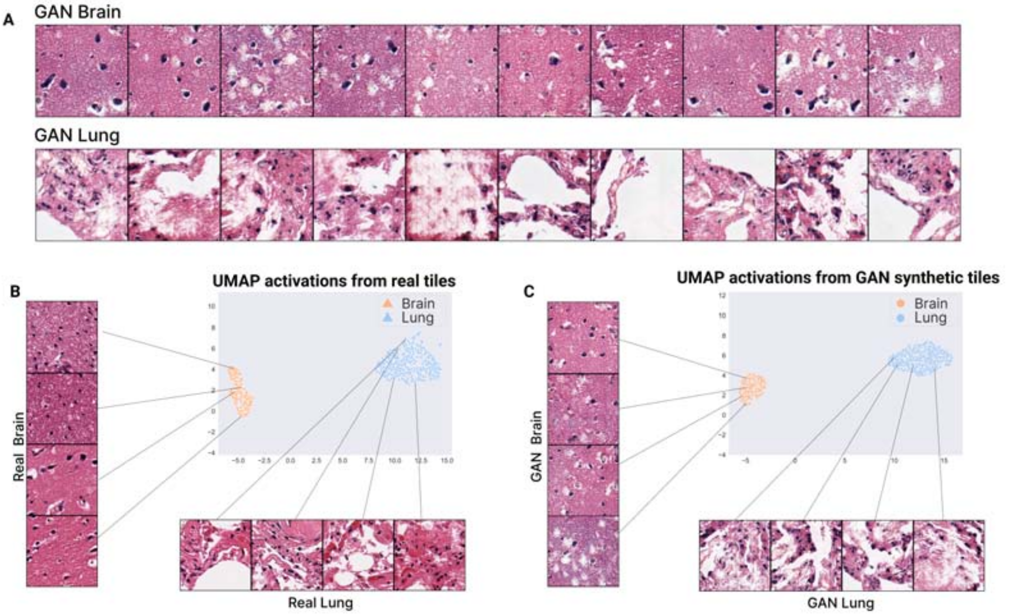


Figure 6.3: A GAN generates realistic lung and brain cortex tiles. Panel A: Tiles generated by the GAN model for brain tissue on the top and for lung tissue on the bottom. Panel B: UMAP representation of the real patients in the lung and brain cortex dataset. Panel C: UMAP representation of generated tiles using the GAN model. 600 tiles are generated per patient, and then used to compute the feature vectors and the UMAP visualization.

Using latent gene expression profiles as input on GANs improves synthetic H&E tiles quality and reduces training time

Next, we used latent gene expression profiles as input instead of random normal distribution for a GAN model generating

WSI tiles. The gene expression was first forwarded through the pretrained βVAE to reduce the dimensionality and encode it in the latent space. Then, that representation plus a narrowed random normal distribution sampled noise was used as input to the generator, which outputs the synthetic tile (Figure 6.1 C). This model shows that synthetic tiles can be generated with fewer artifacts and better quality of the morphological structures (Figure 6.4 A). To demonstrate that the gene expression latent representation provides actual information to generate the tiles, we sampled only from a scaled random normal distribution (values between $[-0.3, 0.3]$) to train the model. This model was not able to produce quality samples of each tissue (Figure 6.4 C).

We also obtained the feature vector from one of the last convolutional layers of the Inception V3 architecture, to observe if the distribution of the synthetic tiles was similar to that one from real patients. The differences between the tissues were preserved as well as the tissue inner-cluster distribution (Figure 6.4 B).

To test the generalization capabilities of the trained models, we also used as input external brain cortex and lung tissue RNA-Seq data. The model was able to successfully generate tissue samples with characteristics similar to those obtained with the training data (Figure 6.4 D). We then tested whether a model trained on real data can distinguish the synthetic generated tiles from this GEO cohort. This model reached an accuracy of 80.5%, a F1-Score of 79.7%, and an AUC of 0.805, showing that a model trained on real tiles can accurately classify the synthetic tiles. Finally, we observed that the RNA expression-infused GAN model needed fewer training epochs in comparison to the regular GAN model (Figure 6.5).

Expert evaluation of synthetic tiles

Next, we asked a panel off five board-certified anatomic pathologists with different subspecialty expertise to rate the quality of brain cortex and lung cortex tiles. These pathologists were not informed about the presence of synthetic data in the examples. The pathologist's evaluation of the morphological structures resulted in a mean score of 3.55 ± 0.95 for real brain, 2.88 ± 0.62 for GAN brain and 2.94 ± 0.64 for RNA-GAN brain (the scores range between 1 and 5). For the lung tissue, the mean score for the real samples was

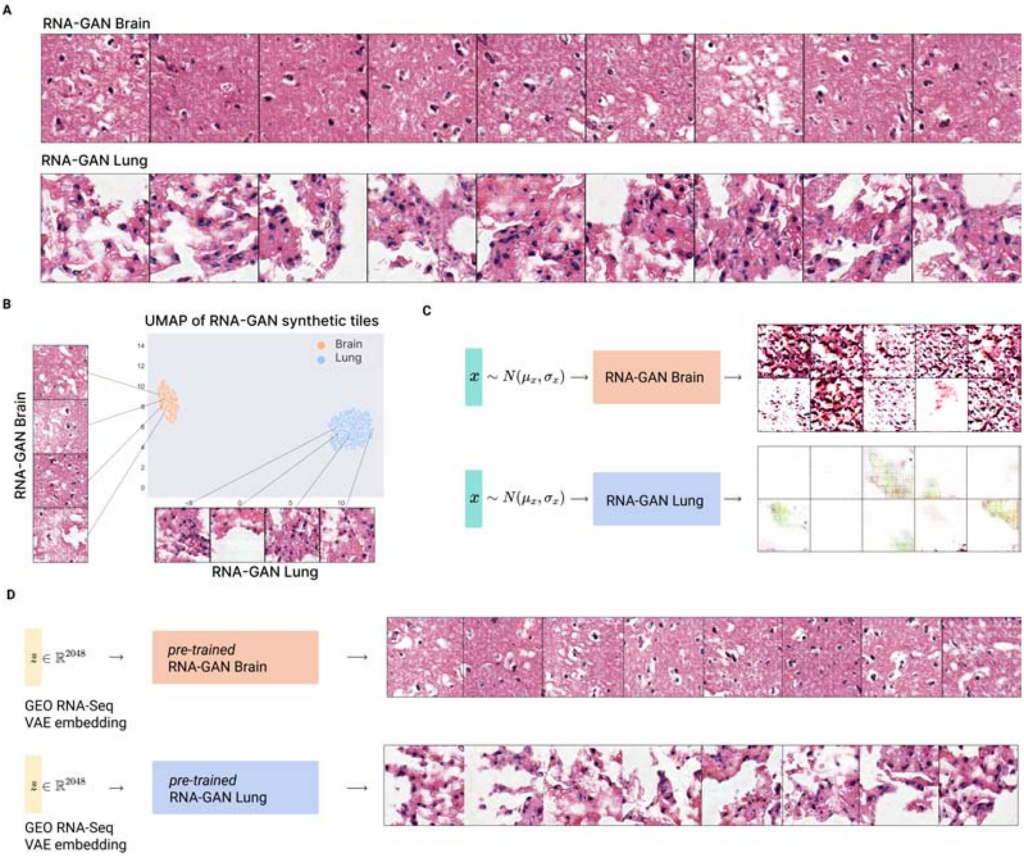


Figure 6.4: A gene expression infused GAN improves-tile quality. Panel A: tiles generated using the RNA-GAN model for lung and brain cortex healthy tissue. Panel B: UMAP visualization of the patients by generating tiles using their gene expression. The model preserves the distribution differences between the two tissues. Panel C: generated tiles of the model trained using only random gaussian data on a small range ($[-0.3, 0.3]$) do not generate high-quality tiles, showing that the gene expression distribution is essential for synthetic tile generation. Panel D: brain cortex and lung tissue tiles generated using an external data set (GSE120795), showing the generalization capabilities of the model.

2.26 ± 1.14, 1 ± 0.55 for the GAN lung, and 1.73 ± 0.79 for the RNA-GAN lung. Significant differences were found in the pathologists evaluation between the GAN and RNA-GAN lung synthetic tiles scores ($p - value = 0.025$), and no significant differences between the real and RNA-GAN lung tiles scores ($p - value = 0.052$). A bigger mean evaluation score difference was found between real and GAN tiles than between real and RNA-GAN tiles, showing that the quality of RNA-GAN synthetic tiles is closer to real tiles (Figure 6.6 A). In addition, the mean difference in the evaluation between GAN and RNA-GAN tiles was bigger than zero, showing the preference of pathologists for the RNA-GAN tiles (Figure

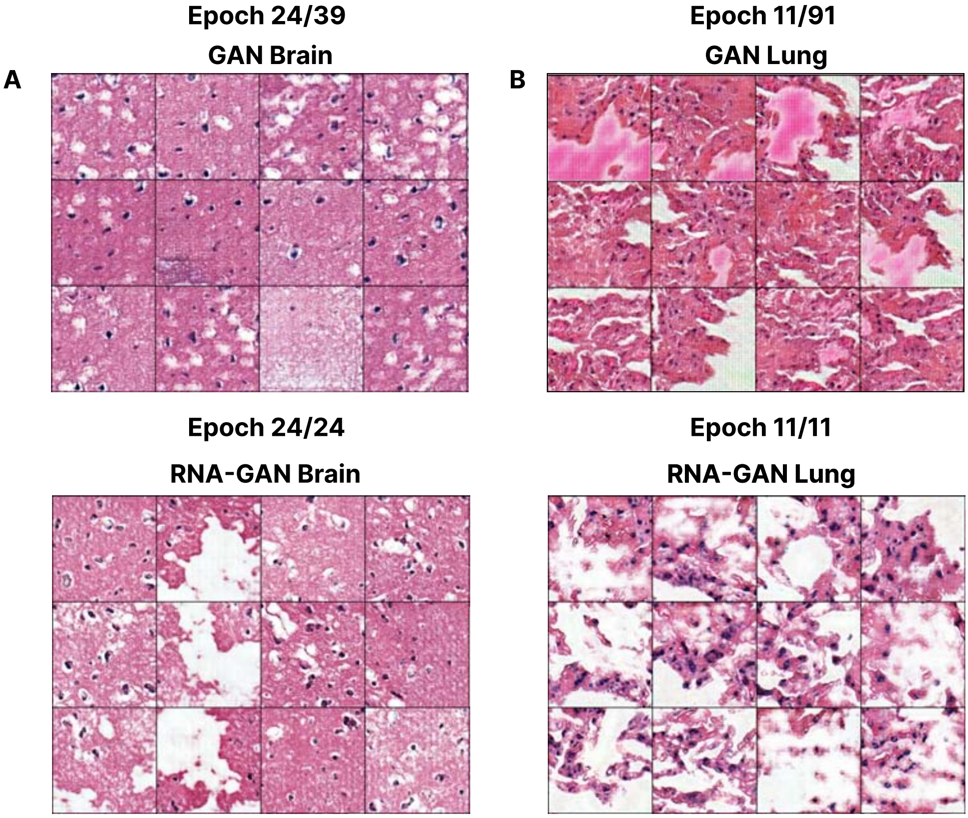


Figure 6.5: A gene expression profile-infused GAN converges faster: brain cortex and lung tissue tiles generated at the same epoch during training for the model with and without gene expression profiles. The visualized epoch is the last epoch of training for the models using RNA-Seq data. Panel A: brain cortex generation at training epoch 24 for GAN and RNA-GAN models, with similar performance and quality between the generated tiles, however, less diversity is obtained when not using gene expression profiles. Panel B: lung tissue generation at training epoch 11 for both the GAN and RNA-GAN models. A comparison of both models show noticeable differences in the quality of the generated tiles. The model using gene expression profiles outputs better morphological features, less artifacts, and has a higher overall quality.

6.6 B). Pathologists detected the tissue of origin of the tiles with a 100% ± 0.0 accuracy for both real and RNA-GAN tiles and with a 74.98% ± 20.40 accuracy for the GAN tiles. In addition, pathologists reported significantly fewer artifacts in RNA-GAN images (56% in comparison to 70% for GAN images).

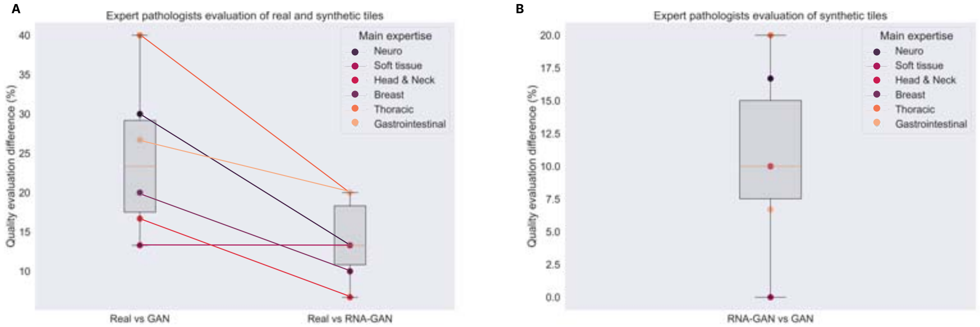


Figure 6.6: Expert evaluation of synthetic slides. Panel A: difference in morphological structure quality of synthetic (generated by GAN and RNA-GAN) and real tissues based on the pathologists evaluation. The difference between real tiles and generated tiles was bigger for GAN than for RNA-GAN. Panel B: Difference in morphological structure quality between the synthetic generated tiles by the GAN and RNA-GAN based on the pathologists evaluation. Pathologists evaluated better those tiles generated using RNA-GAN in comparison to only GAN.

6.5 Conclusions

In this work, we have presented a novel methodology for the RNA-to-image task. We trained a βVAE to reduce the dimensionality of the gene expression data, and obtain a representative latent space. βVAE are able to obtain a latent representation that encodes the characteristics of different healthy tissues (see Figure 6.2 A). Not only that but the generative capabilities of the model are also tested to ensure that the latent space is representative of the gene expression differences(see Figure 6.2 C).

Firstly, we showed how GANs can generate more realistic synthetic tiles when they are infused with the latent gene expression representation, successfully generating healthy lung and brain cortex tissue. In addition, the training time was highly reduced, needing 88% fewer epochs for the RNA-GAN model to generate realistic tiles. This shows how important it is to include information from other modalities, helping to guide the generation process. We showed how our trained model had generalization capabilities, using RNA-Seq data from out of the distribution and generating synthetic WSI tiles. However, GANs have two major drawbacks for the RNA-to-image task. Their training is unstable and prone to model collapse, leading to loss of diversity in the generated samples [100, 246, 247]. Recently presented models, such as diffusion models, can arise as a solution to these problems [229], and their use is something that we would like to explore.

[229]: Saharia et al. (2022), “Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding”

RNA-to-image synthesis can be utterly important for data imputation and data augmentation, and to study the relations between these two modalities. It has been shown that the performance of DNNs increases with the number of samples, and our generated synthetic samples can be used for pre-training purposes. Aside from their practical utility in data augmentation, models such as the RNA-GAN, which utilize latent representations of an entire RNA-Seq profile, might allow for the identification of novel morphologic features associated with clinically relevant molecular biological states that are currently unrecognized by the human eye.

In future work, we will experiment with different architectures to improve sample quality, such as recently presented diffusion models and apply them in a multi-cancer setting to study the capabilities of these models for the RNA-to-Image synthesis task [229].

Final remarks

7 Conclusions & future work

This chapter aims to summarize the outcomes of this thesis, highlighting the most relevant achievements. It further outlines some lines of work that will be pursued next.

7.1 Final conclusions

In this thesis, we have focused on developing multi-modal ML models for developing advanced CDSS in the context of precision medicine for cancer disease. The goals of the thesis (see 3) were accomplish and described in Chapters 4, 5, and 6.

Firstly, we studied how to integrate the information provided by two complementary modalities, WSI, and gene expression for the NSCLC subtype classification problem (see Chapter 4). They describe the same phenomenon (NSCLC subtypes) from two different perspectives and at two different scales. We used the state-of-the-art classification and preprocessing pipelines at the time of the development to create two classification models, and fused the information in a late fusion scheme. These were CNNs for the case of imaging data and SVM+feature selection for genomic data. By fusing the probabilities provided by two classifiers, the fusion model outperformed the single-modality classifiers, showing the capabilities of multi-modal classification models. We also showed how the selected genes had biological relevance, and how they were associated with NSCLC subtypes, validating previous results showing how the use of computational methods such as mRMR can give an insight into the changes in the gene expression of different diseases or states [112, 161, 203].

Encouraged by these results, we focused on increasing the modalities used for the classification (see Chapter 5). Cancer can be described at multiple levels, and the different screenings can reflect that. We incorporated three new modalities, DNA Methylation, miRNA-Seq, and Copy Number Variation. Our goal was to study how these modalities play together for this specific problem, which ones improve the classification performance and how they affect the predictions made by the other modalities. To do so, we created a different fusion model for all possible groupings of the modalities (having two-modality, three-modality, four-modality, and five-modality classifiers). Previously, fusion models that fused the probabilities from different classifiers only used the performance of each classifier to assign a weight to each modality [174–177, 208]. However, by doing so we are not taking into account how well the information provided by the classifiers mixes together. Therefore, we presented a novel scheme to optimize the weights assigned to the different modalities and classes using stochastic gradient descent. Thus, we are also considering how important is that modality with respect to the rest of the fusion. Our final presented model (using all modalities) was capable to deal with missing information and outperform each isolated classifier. In addition, it improved or equals the performance of single-modality models presented in the literature for the same problem.

One main problem during the development of this thesis was the lack of data. Biological data is scarce, and usually very expensive. Not all modalities are always obtained and that limits the potential of training multi-modal models. Therefore, in Chapter 6, we presented a solution for the RNA-to-image synthesis tasks. Gene expression and DNA Methylation prediction from WSI have already been proposed in literature [202, 213], but not the other way around. Therefore, we decided to proposed a new approach for this task. We used GANs infused with the gene expression profiles (so-called RNA-GAN) to generate two different types of healthy tissue. The generated tiles were preferred by pathologists and assigned a better score than synthetic tiles generated with a normal GAN. In addition, the RNA-GAN model converged much faster than the normal GAN model, saving training time.

These three works expand the knowledge of multi-modal classification systems that are able to enhance CDSS in the area of cancer. We showed how multi-modal classification models are more robust than single-modality classifiers, and therefore, they can improve the performance of a diagnosis. Also, we showed how multi-modal generative models can be used as a tool to fight data scarcity and to input modalities that are missing, increasing the size of the datasets and, subsequently, improving the model performance. Next, we are going to describe the future lines of work that are going to be pursued, as well as current challenges that need to be faced.

7.2 Future work

Artificial intelligence is revolutionizing the way we interact and understand biology. New advances are presented at an unprecedented pace, showing the usefulness of these techniques. Alphafold [248], for instance, has impacted protein structure prediction massively, allowing biologists to iterate much faster on their experiments. The potential of ML models in science has just started, serving as a tool for discovery. However, these methods are incredibly data-hungry, and the computational requirements that are demanded can only be fulfilled by a handful of organizations. Nevertheless, the impact that these methods are going to have in the next years, specifically for precision medicine, is going to surpass all our expectations. Data is going to become widely available and increasingly multi-modal, therefore, new methods and studies are necessary.

[248]: Jumper et al. (2021), "Highly accurate protein structure prediction with AlphaFold"

This thesis has presented how multi-modal classifiers improve the performance of ML models over using a single modality. This shows a new way of creating CDSS, using all the data available from the patient to have a better understanding of the phenomena we are modeling. New technologies are emerging, such as spatial transcriptomics, that allow us to obtain a spatial representation of the gene expression across the tissue [249, 250]. Right now the granularity is at multiple cells at the same time (usually between 5 and 10) but we are reaching the level of having a single-cell granularity. This would create tons of possibilities since we will be able to map how the tumor is spreading across the tissue. We would like to apply the generative models presented in this thesis to this type of data, given that we will have a real correspondence between gene expression and tile. Also, to improve the capabilities of the generative models presented in this thesis, we would like to explore the use of recently presented diffusion models for the generation of multi-cancer tiles [226, 229, 233], in order to overcome some of the limitations of GANs.

[249]: Burgess (2019), "Spatial transcriptomics coming of age"
[250]: Rao et al. (2021), "Exploring tissue architecture using spatial transcriptomics"

[226]: Ho et al. (2020), "Denoising diffusion probabilistic models"
[229]: Saharia et al. (2022), "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding"
[233]: Ho et al. (2022), "Cascaded diffusion models for high fidelity image generation"

WSI analysis is also rapidly evolving, and new methods are being used. Graph Neural Networks (GNNs) are rising as a new way of representing the information encoded in the slide. As presented in this thesis, WSIs are too big to be processed at once. However, we can build a graph that encodes the spatial and visual features and learn from them. GNNs are successfully being used in a variety of tasks, such as spatial transcriptomics prediction (in combination with a Transformer) [251], grading colorectal cancer [252], or how they can provide interpretable contextual features of clear cell renal cell carcinoma [253]. Vision Transformers (ViT) [254], an extension of the successful transformer architecture adapted to images [225], have also been proposed as a solution to the spatial representation of the WSI. ViT incorporates more global information than traditional CNNs at lower layers, leading to quantitatively different features [255]. In addition, it has been shown that they are more robust to adversarial attacks, which is desired when dealing with patient data [256]. These improvements, along with the patching methodology that is intrinsic to ViT, have made them really interesting for processing WSI. Recently Chen et al. have proposed a way to scale ViT to WSIs, by grouping the features obtained at different scales [257], showing that ViT paired with a self-supervised methodology, can learn relevant biological features. We would like to explore how this model can be adapted in multi-modal settings, and if they provide more information and increased performance in CDSS.

Self-supervised learning offers a new way of learning important features without having to obtain more data. Quiros et al. presented how using self-supervised learning on lung cancer data could unveil morphological characteristics related to subtypes and prognosis [258]. However, self-supervised learning in multi-modal biomedical data has not been fully explored yet. We would like to explore the possibilities of obtaining aligned representations of different modalities in a self-supervised way, reducing the necessity of huge quantities of data to learn a downstream task.

Finally, we would like to incorporate more modalities into our generative models. By doing so, a richer representation can be obtained and finer control of the morphological features that appear in the tissue may be achieved.

[251]: Zeng et al. (2022), "Spatial transcriptomics prediction from histology jointly through Transformer and graph neural networks"

[252]: Zhou et al. (2019), "Cgc-net: Cell graph convolutional network for grading of colorectal cancer histology images"

[253]: Lee et al. (2022), "Derivation of prognostic contextual histopathological features from whole-slide images of tumours via graph deep learning"

[254]: Dosovitskiy et al. (2020), "An image is worth 16x16 words: Transformers for image recognition at scale"

[225]: Vaswani et al. (2017), "Attention is all you need"

[255]: Raghu et al. (2021), "Do vision transformers see like convolutional neural networks?"

[256]: Ghaffari Laleh et al. (2022), "Adversarial attacks and adversarial robustness in computational pathology"

[257]: Chen et al. (2022), "Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning"

[258]: Quiros et al. (2022), "Self-supervised learning unveils morphological clusters behind lung cancer types and prognosis"

Appendix

Publications

International journal with impact factor included in this thesis

Carrillo-Perez, F., Morales, J. C., Castillo-Secilla, D., Molina-Castro, Y., Guillén, A., Rojas, I., & Herrera, L. J. (2021). Non-small-cell lung cancer classification via RNA-Seq and histology imaging probability fusion. *BMC bioinformatics*, 22(1), 1-19.

Carrillo-Perez, F., Morales, J. C., Castillo-Secilla, D., Gevaert, O., Rojas, I., & Herrera, L. J. (2022). Machine-Learning-Based Late Fusion on Multi-Omics and Multi-Scale Data for Non-Small-Cell Lung Cancer Diagnosis. *Journal of Personalized Medicine*, 12(4), 601.

Other international journals with impact factor

Pérez, M. M., **Carrillo-Perez, F.**, Tejada-Casado, M., Ruiz-López, J., Benavides-Reyes, C., & Herrera, L. J. (2022). CIEDE2000 lightness, chroma and hue human gingiva thresholds. *Journal of Dentistry*, 124, 104213.

Carrillo-Perez, F., Pecho, O. E., Morales, J. C., Paravina, R. D., Della Bona, A., Ghinea, R., ... & Herrera, L. J. (2022). Applications of artificial intelligence in dentistry: A comprehensive review. *Journal of Esthetic and Restorative Dentistry*, 34(1), 259-280.

Durand, L. B., Ruiz-López, J., Perez, B. G., Ionescu, A. M., **Carrillo-Pérez, F.**, Ghinea, R., & Perez, M. M. (2021). Color, lightness, chroma, hue, and translucency adjustment potential of resin composites using CIEDE2000 color difference formula. *Journal of Esthetic and Restorative Dentistry*, 33(6), 836-843.

Carrillo-Perez, F., Herrera, L. J., Carceller, J. M., & Guillén, A. (2021). Deep learning to classify ultra-high-energy cosmic rays

by means of PMT signals. *Neural Computing and Applications*, 33(15), 9153-9169.

Castillo-Secilla, D., Gálvez, J. M., **Carrillo-Perez, F.**, Verona-Almeida, M., Redondo-Sánchez, D., Ortuno, F. M., ... & Rojas, I. (2021). KnowSeq R-Bioc package: The automatic smart gene expression tool for retrieving relevant biological knowledge. *Computers in Biology and Medicine*, 133, 104387.

Medeiros, J. A., Pecho, O. E., Pérez, M. M., **Carrillo-Pérez, F.**, Herrera, L. J., & Della Bona, A. (2021). Influence of background color on color perception in dentistry. *Journal of Dentistry*, 108, 103640.

Molina-Molina, A., Ruiz-Malagón, E. J., **Carrillo-Pérez, F.**, Roche-Seruendo, L. E., Damas, M., Banos, O., & García-Pinillos, F. (2020). Validation of mDurance, a wearable surface electromyography system for muscle activity assessment. *Frontiers in Physiology*, 11, 606287.

Pérez, M. M., Della Bona, A., **Carrillo-Pérez, F.**, Dudea, D., Pecho, O. E., & Herrera, L. J. (2020). Does background color influence visual thresholds?. *Journal of Dentistry*, 102, 103475.

Herrera, L. J., Todero Peixoto, C. J., Baños, O., Carceller, J. M., **Carrillo, F.**, & Guillén, A. (2020). Composition classification of ultra-high energy cosmic rays. *Entropy*, 22(9), 998.

Pérez, M. M., Herrera, L. J., **Carrillo, F.**, Pecho, O. E., Dudea, D., Gasparik, C., ... & Della Bona, A. (2019). Whiteness difference thresholds in dentistry. *Dental Materials*, 35(2), 292-297.

Perez, M. M., Ghinea, R., Herrera, L. J., **Carrillo, F.**, Ionescu, A. M.,& Paravina, R. D. (2018). Color difference thresholds for computer-simulated human Gingiva. *Journal of Esthetic and Restorative Dentistry*, 30(2), E24-E30.

International conferences

Carrillo-Perez, F., Morales, J. C., Castillo-Secilla, D., Guillen, A., Rojas, I., & Herrera, L. J. (2021, July). Comparison of fusion methodologies using CNV and RNA-seq for cancer classification: a case study on non-small-cell lung cancer. *In International Conference on Bioengineering and Biomedical Signal and Image Processing* (pp. 339-349). Springer, Cham.

Toledano Pavón, J., Morales Vega, J. C., **Carrillo-Perez, F.**, Herrera, L. J., & Rojas, I. (2021, July). COVID-19 Detection Method from Chest CT Scans via the Fusion of Slice Information and Lung Segmentation. *In International Conference on Bioengineering and Biomedical Signal and Image Processing* (pp. 155-165). Springer, Cham.

Morales Vega, J. C., **Carrillo-Perez, F.**, Toledano Pavón, J., Herrera Maldonado, L. J., & Rojas Ruiz, I. (2021, June). Ensemble Models for Covid Prediction in X-Ray Images. *In International Work-Conference on Artificial Neural Networks* (pp. 559-569). Springer, Cham.

Morales, J. C., **Carrillo-Perez, F.**, Castillo-Secilla, D., Rojas, I., & Herrera, L. J. (2020, May). Enhancing breast cancer classification via information and multi-model integration. *In International Work-Conference on Bioinformatics and Biomedical Engineering* (pp. 750-760). Springer, Cham.

Carrillo-Perez, F., Herrera, L. J., Carceller, J. M., & Guillén, A. (2019, June). Improving classification of ultra-high energy cosmic rays using spacial locality by means of a convolutional DNN. *In International Work-Conference on Artificial Neural Networks* (pp. 222-232). Springer, Cham.

Carrillo-Perez, F., Diaz-Reyes, I., Damas, M., Banos, O., Soto-Hermoso, V. M., & Molina-Molina, A. (2018, September). A novel automated algorithm for computing lumbar flexion test ratios enhancing athletes objective assessment of low back pain. *In Proceedings of the 6th International Congress on Sport Sciences Research and Technology Support* (Seville: SCITEPRESS–Science and Technology Publications) (pp. 34-39).

Grants and awards

Grants

This thesis has been supported by the following projects and grants. The funders had no role in the study design, data collection, and analysis, decision to publish, or preparation of the manuscripts that supported this doctoral thesis:

1. National Research Project **RTI2018-101674-B-I00**, funded by the Spanish Ministry of Sciences, Innovation and Universities.
2. National Research Project **PID2021-128317OB-I00**, funded by the Spanish Ministry of Sciences, Innovation and Universities.
3. Regional Research Project **P20-00163**, funded by Junta de Andalucia.
4. Predoctoral Fulbright Scholarship funded by the Spanish Fulbright commission.

Awards

1. Recipient of one of the 18 predoctoral research Fulbright Spain scholarship for doing a eleven month research stay during 2021-2022 at Stanford University, California, under the supervision of Prof. Olivier Gevaert.

Bibliography

[1] Sara Stinson, Barry Bogin, and Dennis H O'Rourke. *Human biology: an evolutionary and biocultural perspective*. John Wiley & Sons, 2012.

[2] Michal Krassowski, Vivek Das, Sangram K Sahu, and Biswapriya B Misra. "State of the field in multi-omics research: From computational needs to data mining and sharing". In: *Frontiers in Genetics* 11 (2020), p. 610798.

[3] *List of omics topics in biology*. June 2022. URL: `https://en.wikipedia.org/wiki/List%5C_of%5C_omics_topics%5C_in_biology`.

[4] Hyuna Sung et al. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: a cancer journal for clinicians* 71.3 (2021), pp. 209–249.

[5] Zohar Yakhini and Igor Jurisica. *Cancer computational biology*. 2011.

[6] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. "Multi-omics approaches to disease". In: *Genome biology* 18.1 (2017), pp. 1–15.

[7] Richard Hodson. "Precision medicine". In: *Nature* 537.7619 (2016), S49–S49.

[8] Issam El Naqa and Martin J Murphy. "What is machine learning?" In: *machine learning in radiation oncology*. Springer, 2015, pp. 3–11.

[9] Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. "Machine learning applications in cancer prognosis and prediction". In: *Computational and structural biotechnology journal* 13 (2015), pp. 8–17.

[10] Fred Bunz. *Principles of cancer genetics*. Vol. 1. Springer, 2008.

[11] James D Watson and Francis HC Crick. "Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid". In: *Nature* 171.4356 (1953), pp. 737–738.

[12] Anne Sayre. *Rosalind Franklin and dna*. WW Norton & Company, 2000.

[13] Nigel Chaffey. *Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. Molecular biology of the cell. 4th edn.* 2003.

[14] Zephyris. *The structure of the DNA double helix.* June 2022. URL: `https://commons.wikimedia.org/w/index.php?curid=15027555`.

[15] Sydney Brenner, Jeffrey H Miller, and William Broughton. *Encyclopedia of genetics*. Sirsi) i9780122270802. 2002.

[16] National Human Genome Research Institute. *Chromosome.* June 2022. URL: `https://www.genome.gov/genetics-glossary/Chromosome`.

[17] National Institute of General Medical Sciences. *What is genetics?* June 2022. URL: `https://www.nigms.nih.gov/education/fact-sheets/Pages/genetics.aspx`.

[18] Bruce Alberts. *Molecular biology of the cell*. WW Norton & Company, 2017.

[19] Geoffrey M Cooper, Robert E Hausman, and Robert E Hausman. *The cell: a molecular approach*. Vol. 4. ASM press Washington, DC, 2007.

[20] Lauren Pecorino. *Molecular biology of cancer: mechanisms, targets, and therapeutics*. Oxford university press, 2021.

[21] National Cancer Institute. *What is cancer?* June 2022. URL: `https://www.cancer.gov/about-cancer/understanding/what-is-cancer`.

[22] Andrea Ferretti. *Cancer vs Normal Ceell Division*. June 2022. URL: `https://www.kindpng.com/imgv/hbwmixJ_picture-cancer-vs-normal-cell-division-hd-png/`.

[23] Raymond W Ruddon. *Cancer biology*. Oxford University Press, 2007.

[24] Christopher Mader. "The biology of cancer". In: *The Yale Journal of Biology and Medicine* 80.2 (2007), p. 91.

[25] J Craig Venter et al. "The sequence of the human genome". In: *science* 291.5507 (2001), pp. 1304–1351.

[26] Sergey Nurk et al. "The complete sequence of a human genome". In: *Science* 376.6588 (2022), pp. 44–53.

[27] Christine M Micheel, Sharly J Nass, Gilbert S Omenn, et al. "Omics-based clinical discovery: Science, technology, and applications". In: *Evolution of Translational Omics: Lessons Learned and the Path Forward*. National Academies Press (US), 2012.

[28] David P Clark and Nanette J Pazdernik. *Molecular biology*. Elsevier, 2013.

[29] Matthias Meyer and Martin Kircher. "Illumina sequencing library preparation for highly multiplexed target capture and sequencing". In: *Cold Spring Harbor Protocols* 2010.6 (2010), pdb–prot5448.

[30] Sam Behjati and Patrick S Tarpey. "What is next generation sequencing?" In: *Archives of Disease in Childhood-Education and Practice* 98.6 (2013), pp. 236–238.

[31] Aimaiti Yasen et al. "Progress and applications of single-cell sequencing techniques". In: *Infection, Genetics and Evolution* 80 (2020), p. 104198.

[32] Yukie Kashima et al. "Single-cell sequencing techniques from individual to multiomics analyses". In: *Experimental & Molecular Medicine* 52.9 (2020), pp. 1419–1427.

[33] Rory Stark, Marta Grzelak, and James Hadfield. "RNA sequencing: the teenage years". In: *Nature Reviews Genetics* 20.11 (2019), pp. 631–656.

[34] Daniel Castillo Secilla et al. "Integration of heterogeneous gene expression sources in human cancer pathologies, employing high performance computing and machine learning techniques". In: (2020).

[35] DMLapato. *Illumina dye sequencing*. June 2022. URL: `https://en.wikipedia.org/wiki/File:Cluster_Generation.png#/media/File:Cluster_Generation.png`.

[36] Ralph A Bradshaw and Philip D Stahl. *Encyclopedia of cell biology*. Academic Press, 2015.

[37] Geraldo A Passos. *Transcriptomics in Health and Disease*. Springer, 2015.

[38] Jiaqian Wu and Dong Kim. *Transcriptomics and gene regulation*. Springer, 2016.

[39] Malachi Griffith, Jason R Walker, Nicholas C Spies, Benjamin J Ainscough, and Obi L Griffith. "Informatics for RNA sequencing: a web resource for analysis on the cloud". In: *PLoS computational biology* 11.8 (2015), e1004393.

[40] Zhen He, Rong Zhang, Feng Jiang, Wenjing Hou, and Cheng Hu. "Role of genetic and environmental factors in DNA methylation of lipid metabolism". In: *Genes & diseases* 5.1 (2018), pp. 9–15.

[41] Xu Gao, Yan Zhang, Lutz Philipp Breitling, and Hermann Brenner. "Tobacco smoking and methylation of genes related to lung cancer development". In: *Oncotarget* 7.37 (2016), p. 59017.

[42] Allen S Yang et al. "A simple method for estimating global DNA methylation using bisulfite PCR of repetitive DNA elements". In: *Nucleic acids research* 32.3 (2004), e38–e38.

[43] Azim Surani. *Epigenomics: From chromatin biology to therapeutics*. Cambridge University Press, 2012.

[44] Mark A Dawson and Tony Kouzarides. "Cancer epigenetics: from mechanism to therapy". In: *cell* 150.1 (2012), pp. 12–27.

[45] Liron Pantanowitz et al. "Twenty years of digital pathology: an overview of the road travelled, what is on the horizon, and the emergence of vendor-neutral archives". In: *Journal of pathology informatics* 9.1 (2018), p. 40.

[46] Matthew G Hanna, Anil Parwani, and Sahussapont Joseph Sirintrapun. "Whole slide imaging: technology and applications". In: *Advances in Anatomic Pathology* 27.4 (2020), pp. 251–259.

[47] Liron Pantanowitz et al. "Review of the current state of whole slide imaging in pathology". In: *Journal of pathology informatics* 2.1 (2011), p. 36.

[48] Robert W Ogilvie. *Virtual microscopy and virtual slides in teaching, diagnosis, and research*. CRC Press, 2005.

[49] Navid Farahani, Anil V Parwani, Liron Pantanowitz, et al. "Whole slide imaging in pathology: advantages, limitations, and emerging perspectives". In: *Pathol Lab Med Int* 7.23-33 (2015), p. 4321.

[50] Andrew J Evans, Mohamed E Salama, Walter H Henricks, and Liron Pantanowitz. "Implementation of whole slide imaging for clinical purposes: issues to consider from the perspective of early adopters". In: *Archives of pathology & laboratory medicine* 141.7 (2017), pp. 944–959.

[51] John KC Chan. "The wonderful colors of the hematoxylin–eosin stain in diagnostic surgical pathology". In: *International journal of surgical pathology* 22.1 (2014), pp. 12–32.

[52] Shaimaa Al-Janabi, André Huisman, and Paul J Van Diest. "Digital pathology: current status and future perspectives". In: *Histopathology* 61.1 (2012), pp. 1–9.

[53] Hamid Reza Tizhoosh and Liron Pantanowitz. "Artificial intelligence and digital pathology: challenges and opportunities". In: *Journal of pathology informatics* 9.1 (2018), p. 38.

[54] Matthew G Hanna et al. “Integrating digital pathology into clinical practice”. In: *Modern Pathology* 35.2 (2022), pp. 152–164.

[55] Yinhai Wang, Kate E Williamson, Paul J Kelly, Jacqueline A James, and Peter W Hamilton. “SurfaceSlide: a multitouch digital pathology platform”. In: *PloS one* 7.1 (2012), e30783.

[56] Librepath. *Micrograph showing optic nerve head and retina, H&E stain.* June 2022. URL: `https://commons.wikimedia.org/wiki/File:Retina%5C_--%5C_high_mag.jpg`.

[57] John N Weinstein et al. “The cancer genome atlas pan-cancer analysis project”. In: *Nature genetics* 45.10 (2013), p. 1113.

[58] John Lonsdale et al. “The genotype-tissue expression (GTEx) project”. In: *Nature genetics* 45.6 (2013), pp. 580–585.

[59] Kenneth Clark et al. “The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository”. In: *Journal of digital imaging* 26.6 (2013), pp. 1045–1057.

[60] World Health Organization. *World Cancer Report 2014*. 2014.

[61] Cancer Research UK. *Types of lung cancer*. `https://www.cancerresearchuk.org/about-cancer/lung-cancer/stages-types-grades/types`.

[62] Peter Goldstraw et al. “Non-small-cell lung cancer”. In: *The Lancet* 378.9804 (2011), pp. 1727–1740.

[63] Janakiraman Subramanian and Ramaswamy Govindan. “Lung cancer in never smokers: a review”. In: *Journal of clinical oncology* 25.5 (2007), pp. 561–570.

[64] Stacey A Kenfield, Esther K Wei, Meir J Stampfer, Bernard A Rosner, and Graham A Colditz. “Comparison of aspects of smoking among the four histological types of lung cancer”. In: *Tobacco control* 17.3 (2008), pp. 198–204.

[65] William D. Travis, Lois B. Travis, and Susan S. Devesa. “Lung cancer”. In: *Cancer* 75.S1 (1995), pp. 191–202.

[66] Mary Gospodarowicz and Brian O’Sullivan. “Prognostic factors in cancer”. In: *Seminars in surgical oncology*. Vol. 21. 1. Wiley Online Library. 2003, pp. 13–18.

[67] Gilbert Massard, Gaetano Rocco, and Federico Venuta. “The European educational platform on thoracic surgery”. In: *Journal of Thoracic Disease* 6.Suppl 2 (2014), S276.

[68] William D Travis. “Pathology of lung cancer”. In: *Clinics in chest medicine* 23.1 (2002), pp. 65–81.

[69] Cesare Gridelli et al. “Non-small-cell lung cancer”. In: *Nature reviews Disease primers* 1.1 (2015), pp. 1–16.

[70] David S Ettinger et al. “Non–small cell lung cancer”. In: *Journal of the national comprehensive cancer network* 8.7 (2010), pp. 740–801.

[71] David S Ettinger et al. “Non–small cell lung cancer”. In: *Journal of the National Comprehensive Cancer Network* 10.10 (2012), pp. 1236–1271.

[72] Konstantinos Zarogoulidis et al. "Treatment of non-small cell lung cancer (NSCLC)". In: *Journal of thoracic disease* 5.Suppl 4 (2013), S389.

[73] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. 4. Springer, 2006.

[74] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[75] Stuart J Russell and Peter Norvig. *Artificial intelligence a modern approach*. Pearson Education, Inc., 2010.

[76] Martijn van Otterlo and Marco Wiering. "Reinforcement learning and markov decision processes". In: *Reinforcement learning*. Springer, 2012, pp. 3–42.

[77] David Silver et al. "Mastering the game of Go with deep neural networks and tree search". In: *nature* 529.7587 (2016), pp. 484–489.

[78] David Silver et al. "Mastering the game of go without human knowledge". In: *nature* 550.7676 (2017), pp. 354–359.

[79] Oriol Vinyals et al. "Grandmaster level in StarCraft II using multi-agent reinforcement learning". In: *Nature* 575.7782 (2019), pp. 350–354.

[80] Christian Robert. *Machine learning, a probabilistic perspective*. 2014.

[81] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[82] Jürgen Schmidhuber. "Deep learning in neural networks: An overview". In: *Neural networks* 61 (2015), pp. 85–117.

[83] David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. "Learning representations by back-propagating errors". In: *Cognitive modeling* 5.3 (1988), pp. 533–536.

[84] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. "Multilayer feedforward networks are universal approximators". In: *Neural networks* 2.5 (1989), pp. 359–366.

[85] Francisco Carrillo-Perez et al. "Applications of artificial intelligence in dentistry: A comprehensive review". In: *Journal of Esthetic and Restorative Dentistry* 34.1 (2022), pp. 259–280.

[86] Kunihiko Fukushima. "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position". In: *Biological Cybernetics* 36.4 (1980), pp. 193–202.

[87] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[88] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444.

[89] Christopher G Harris, Mike Stephens, et al. "A combined corner and edge detector." In: *Alvey vision conference*. Vol. 15. 50. Citeseer. 1988, pp. 10–5244.

[90] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[91] Saleh Albelwi and Ausif Mahmood. “A framework for designing the architectures of deep convolutional neural networks”. In: *Entropy* 19.6 (2017), p. 242.

[92] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. “Rethinking the inception architecture for computer vision”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826.

[93] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[94] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint* (2014), p. 1409.1556.

[95] Jie Lu et al. “Transfer learning using computational intelligence: A survey”. In: *Knowledge-Based Systems* 80 (2015), pp. 14–23.

[96] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. “A survey of transfer learning”. In: *Journal of Big Data* 3.1 (2016), p. 9.

[97] Qiang Yang, Yu Zhang, Wenyuan Dai, and Sinno Jialin Pan. *Transfer learning*. 1st ed. Cambridge University Press, 2020.

[98] Ian Goodfellow et al. “Generative adversarial nets”. In: *Advances in neural information processing systems* 27 (2014).

[99] Martin Arjovsky, Soumith Chintala, and Léon Bottou. “Wasserstein generative adversarial networks”. In: *International conference on machine learning*. PMLR. 2017, pp. 214–223.

[100] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. “Improved training of wasserstein gans”. In: *Advances in neural information processing systems* 30 (2017).

[101] Alec Radford, Luke Metz, and Soumith Chintala. “Unsupervised representation learning with deep convolutional generative adversarial networks”. In: *arXiv preprint arXiv:1511.06434* (2015).

[102] Andrew Brock, Jeff Donahue, and Karen Simonyan. “Large scale GAN training for high fidelity natural image synthesis”. In: *arXiv preprint arXiv:1809.11096* (2018).

[103] Tero Karras et al. “Alias-free generative adversarial networks”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 852–863.

[104] Biting Yu et al. “Ea-GANs: edge-aware generative adversarial networks for cross-modality MR image synthesis”. In: *IEEE transactions on medical imaging* 38.7 (2019), pp. 1750–1762.

[105] Adalberto Claudio Quiros, Roderick Murray-Smith, and Ke Yuan. “PathologyGAN: Learning deep representations of cancer tissue”. In: *arXiv preprint arXiv:1907.02644* (2019).

[106] Diederik P Kingma and Max Welling. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013).

[107] Irina Higgins et al. “beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework”. In: *ICLR*. 2017.

[108] Corinna Cortes and Vladimir Vapnik. “Support-vector networks”. In: *Machine learning* 20.3 (1995), pp. 273–297.

[109] William S Noble. “What is a support vector machine?” In: *Nature biotechnology* 24.12 (2006), pp. 1565–1567.

[110] Yuanqing Lin et al. “Large-scale image classification: Fast feature extraction and SVM training”. In: *CVPR 2011*. IEEE. 2011, pp. 1689–1696.

[111] Christina Leslie, Eleazar Eskin, and William Stafford Noble. “The spectrum kernel: A string kernel for SVM protein classification”. In: *Biocomputing 2002*. World Scientific, 2001, pp. 564–575.

[112] Daniel Castillo et al. “Leukemia multiclass assessment and classification from Microarray and RNA-seq technologies integration at gene expression level”. In: *PloS one* 14.2 (2019), e0212127.

[113] Anneleen Daemen et al. “A kernel-based integration of genome-wide data for clinical decision support”. In: *Genome medicine* 1.4 (2009), pp. 1–17.

[114] Olivier Gevaert, Frank De Smet, Dirk Timmerman, Yves Moreau, and Bart De Moor. “Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks”. In: *Bioinformatics* 22.14 (2006), e184–e190.

[115] Anika Cheerla and Olivier Gevaert. “Deep learning with multimodal representation for pancancer prognosis prediction”. In: *Bioinformatics* 35.14 (2019), pp. i446–i454.

[116] Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P Lungren. “Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines”. In: *NPJ digital medicine* 3.1 (2020), pp. 1–9.

[117] Gyanendra K Verma and Uma Shanker Tiwary. “Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals”. In: *NeuroImage* 102 (2014), pp. 162–172.

[118] Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders. “Early versus late fusion in semantic video analysis”. In: *Proceedings of the 13th annual ACM international conference on Multimedia*. 2005, pp. 399–402.

[119] Nikhil Cheerla and Olivier Gevaert. “MicroRNA based pan-cancer diagnosis and treatment recommendation”. In: *BMC bioinformatics* 18.1 (2017), pp. 1–11.

[120] Sivaramakrishnan Rajaraman et al. “Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images”. In: *PeerJ* 6 (2018), e4568.

[121] Pablo Garcıéa-Risueño and Pablo E Ibáñez. “A review of High Performance Computing foundations for scientists”. In: *International journal of modern physics C* 23.07 (2012), p. 1230001.

[122] Adam Paszke et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8024–8035.

[123] Stanford Research Computing Center (SRCC). *Sherlock documentation*. Sept. 2022. URL: `https://www.sherlock.stanford.edu/`.

[124] Slurm Team. *Slurm Workload Manager documentation*. Sept. 2022. URL: `https://slurm.schedmd.com/documentation.html`.

[125] Thomas Wolf. *Training Neural Nets on Larger Batches: Practical Tips for 1-GPU, Multi-GPU & Distributed setups*. June 2022. URL: `https://medium.com/huggingface/training-larger-batches-practical-tips-on-1-gpu-multi-gpu-distributed-setups-ec88c3e51255`.

[126] Stanford Research Computing Center (SRCC). *Sherlock computation documentation*. Sept. 2022. URL: `https://www.sherlock.stanford.edu/docs/tech/#computing`.

[127] Nasser Hanna et al. “Systemic therapy for stage IV non–small-cell lung cancer: American Society of Clinical Oncology clinical practice guideline update”. In: *Journal of Clinical Oncology* (2017).

[128] Baoshan Ma, Yao Geng, Fanyu Meng, Ge Yan, and Fengju Song. “Identification of a sixteen-gene prognostic biomarker for lung adenocarcinoma using a machine learning method”. In: *Journal of Cancer* 11.5 (2020), p. 1288.

[129] Diego Ardila et al. “End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography”. In: *Nature medicine* 25.6 (2019), pp. 954–961.

[130] Yutong Xie et al. “Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest CT”. In: *IEEE transactions on medical imaging* 38.4 (2018), pp. 991–1004.

[131] Nicolas Coudray et al. “Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning”. In: *Nature medicine* 24.10 (2018), pp. 1559–1567.

[132] Hoa Hoang Ngoc Pham et al. “Detection of lung cancer lymph node metastases from Whole-Slide histopathologic images using a two-step deep learning approach”. In: *The American journal of pathology* 189.12 (2019), pp. 2428–2439.

[133] Bethany Jill Williams, David Bottoms, and Darren Treanor. “Future-proofing pathology: the case for clinical adoption of digital pathology”. In: *Journal of clinical pathology* 70.12 (2017), pp. 1010–1018.

[134] Andreas Heindl, Sidra Nawaz, and Yinyin Yuan. “Mapping spatial heterogeneity in the tumor microenvironment: a new era for digital pathology”. In: *Laboratory investigation* 95.4 (2015), pp. 377–384.

[135] Jun Cheng et al. “Identification of topological features in renal tumor microenvironment associated with patient survival”. In: *Bioinformatics* 34.6 (2018), pp. 1024–1030.

[136] Nikola Simidjievski et al. “Variational autoencoders for cancer data integration: design principles and computational practice”. In: *Frontiers in genetics* 10 (2019), p. 1205.

[137] Tianle Ma and Aidong Zhang. “Integrate multi-omics data with biological interaction networks using Multi-view Factorization AutoEncoder (MAE)”. In: *BMC genomics* 20.11 (2019), pp. 1–11.

[138] Garam Lee, Byungkon Kang, Kwangsik Nho, Kyung-Ah Sohn, and Dokyoon Kim. “MildInt: deep learning-based multimodal longitudinal data integration framework”. In: *Frontiers in Genetics* 10 (2019), p. 617.

[139] Kyunghyun Cho et al. “Learning phrase representations using RNN encoder-decoder for statistical machine translation”. In: *arXiv preprint arXiv:1406.1078* (2014).

[140] Yu-Heng Lai et al. “overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning”. In: *Scientific reports* 10.1 (2020), pp. 1–11.

[141] Luıés A Vale Silva and Karl Rohr. “Pan-Cancer Prognosis Prediction Using Multimodal Deep Learning”. In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2020, pp. 568–571.

[142] Richard J Chen et al. “Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis”. In: *arXiv preprint arXiv:1912.08937* (2019).

[143] Gonzalo Gómez-López, Joaquıén Dopazo, Juan C Cigudosa, Alfonso Valencia, and Fátima Al-Shahrour. “Precision medicine needs pioneering clinical bioinformaticians”. In: *Briefings in bioinformatics* 20.3 (2019), pp. 752–766.

[144] Johannes Smolander, Alexey Stupnikov, Galina Glazko, Matthias Dehmer, and Frank Emmert-Streib. “Comparing biological information contained in mRNA and non-coding RNAs for classification of lung cancer patients”. In: *BMC cancer* 19.1 (2019), p. 1176.

[145] Zhirui Fan et al. “Identification of an early diagnostic biomarker of lung adenocarcinoma based on co-expression similarity and construction of a diagnostic model”. In: *Journal of translational medicine* 16.1 (2018), p. 205.

[146] Jingming Zhao et al. “Construction of a specific SVM classifier and identification of molecular markers for lung adenocarcinoma based on lncRNA-miRNA-mRNA network”. In: *OncoTargets and therapy* 11 (2018), p. 3129.

[147] Sara González, Daniel Castillo, Juan Manuel Galvez, Ignacio Rojas, and Luis Javier Herrera. “Feature Selection and Assessment of Lung Cancer Sub-types by Applying Predictive Models”. In: *International Work-Conference on Artificial Neural Networks*. Springer. 2019, pp. 883–894.

[148] Mila Efimenko, Alexander Ignatev, and Konstantin Koshechkin. “Review of medical image recognition technologies to detect melanomas using neural networks”. In: *BMC bioinformatics* 21.11 (2020), pp. 1–7.

[149] Fahdi Kanavati et al. “Weakly-supervised learning for lung carcinoma classification using deep learning”. In: *Scientific Reports* 10.1 (), pp. 1–11.

[150] Simon Graham et al. “Classification of lung cancer histology images using patch-level summary statistics”. In: *Medical Imaging 2018: Digital Pathology*. Vol. 10581. International Society for Optics and Photonics. 2018, p. 1058119.

[151] Zhang Li et al. “Computer-aided diagnosis of lung carcinoma using deep learning-a pilot study”. In: *arXiv preprint arXiv:1803.05471* (2018).

[152] Kun-Hsing Yu et al. “Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features”. In: *Nature communications* 7.1 (2016), pp. 1–10.

[153] Pegah Khosravi, Ehsan Kazemi, Marcin Imielinski, Olivier Elemento, and Iman Hajirasouliha. “Deep convolutional neural networks enable discrimination of heterogeneous digital pathology images”. In: *EBioMedicine* 27 (2018), pp. 317–328.

[154] Robert L Grossman et al. “Toward a shared vision for cancer genomic data”. In: *New England Journal of Medicine* 375.12 (2016), pp. 1109–1112.

[155] Adam Goode, Benjamin Gilbert, Jan Harkes, Drazen Jukic, and Mahadev Satyanarayanan. “OpenSlide: A vendor-neutral software foundation for digital pathology”. In: *Journal of pathology informatics* 4 (2013).

[156] *GDC RNA-Seq Analysis Pipeline*. `https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/`. Accessed: 2021-07-06.

[157] Daniel Castillo-Secilla et al. *KnowSeq: A R package to extract knowledge by using RNA-seq raw files*. R package version 1.3.0. 2020.

[158] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[159] Matthew E Ritchie et al. “limma powers differential expression analyses for RNA-sequencing and microarray studies”. In: *Nucleic acids research* 43.7 (2015), e47–e47.

[160] Hanchuan Peng, Fuhui Long, and Chris Ding. “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy”. In: *IEEE Transactions on pattern analysis and machine intelligence* 27.8 (2005), pp. 1226–1238.

[161] Daniel Castillo et al. “Integration of RNA-Seq data with heterogeneous microarray data for breast cancer profiling”. In: *BMC bioinformatics* 18.1 (2017), p. 506.

[162] Juan M Gálvez et al. “Towards improving skin cancer diagnosis by integrating microarray and RNA-seq datasets”. In: *IEEE journal of biomedical and health informatics* 24.7 (2019), pp. 2119–2130.

[163] Chris Ding and Hanchuan Peng. “Minimum redundancy feature selection from microarray gene expression data”. In: *Journal of bioinformatics and computational biology* 3.02 (2005), pp. 185–205.

[164] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).

[165] Mingxing Tan and Quoc V Le. “Efficientnet: Rethinking model scaling for convolutional neural networks”. In: *arXiv preprint arXiv:1905.11946* (2019).

[166] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[167] Olga Russakovsky et al. “Imagenet large scale visual recognition challenge”. In: *International journal of computer vision* 115.3 (2015), pp. 211–252.

[168] Qing Wen, Chang-Sik Kim, Peter W Hamilton, and Shu-Dong Zhang. “A gene-signature progression approach to identifying candidate small-molecule cancer therapeutics with connectivity mapping”. In: *BMC bioinformatics* 17.1 (2016), pp. 1–11.

[169] A Laganà et al. “Integrative network analysis identifies novel drivers of pathogenesis and progression in newly diagnosed multiple myeloma”. In: *Leukemia* 32.1 (2018), pp. 120–130.

[170] Philip S Bernard and Carl T Wittwer. “Real-time PCR technology for cancer diagnostics”. In: *Clinical chemistry* 48.8 (2002), pp. 1178–1185.

[171] Alexandros Kalousis, Julien Prados, and Melanie Hilario. “Stability of feature selection algorithms: a study on high-dimensional spaces”. In: *Knowl Inf Syst* 12.1 (2007), pp. 95–116.

[172] S Sathiya Keerthi and Chih-Jen Lin. “Asymptotic behaviors of support vector machines with Gaussian kernel”. In: *Neural computation* 15.7 (2003), pp. 1667–1689.

[173] Ting-Fan Wu, Chih-Jen Lin, and Ruby C Weng. “Probability estimates for multi-class classification by pairwise coupling”. In: *Journal of Machine Learning Research* 5.Aug (2004), pp. 975–1005.

[174] Yunyun Dong et al. “MLW-gcForest: a multi-weighted gcForest model towards the staging of lung adenocarcinoma based on multi-modal genetic data”. In: *BMC bioinformatics* 20.1 (2019), pp. 1–14.

[175] Tao Meng, Lin Lin, Mei-Ling Shyu, and Shu-Ching Chen. “Histology image classification using supervised classification and multimodal fusion”. In: *2010 IEEE international symposium on multimedia*. IEEE. 2010, pp. 145–152.

[176] Vo Hoang Trong, Yu Gwang-hyun, Dang Thanh Vu, and Kim Jin-young. “Late fusion of multimodal deep neural networks for weeds classification”. In: *Computers and Electronics in Agriculture* 175 (2020), p. 105506.

[177] Adrien Depeursinge et al. “Fusing visual and clinical information for lung tissue classification in high-resolution computed tomography”. In: *Artificial intelligence in medicine* 50.1 (2010), pp. 13–21.

[178] Damien François, Fabrice Rossi, Vincent Wertz, and Michel Verleysen. “Resampling methods for parameter-free and robust feature selection with mutual information”. In: *Neurocomputing* 70.7-9 (2007), pp. 1276–1288.

[179] Denise Carvalho-Silva et al. “Open Targets Platform: new developments and updates two years on”. In: *Nucleic acids research* 47.D1 (2019), pp. D1056–D1065.

[180] Zhibo Tan, Chao Yang, Xiaohan Zhang, Pingju Zheng, and Weixi Shen. “Expression of glucose transporter 1 and prognosis in non-small cell lung cancer: a pooled analysis of 1665 patients”. In: *Oncotarget* 8.37 (2017), p. 60954.

[181] Elisabeth Smolle et al. “Distribution and prognostic significance of gluconeogenesis and glycolysis in lung cancer”. In: *Molecular oncology* 14.11 (2020), pp. 2853–2867.

[182] Huanyu Zhao et al. “Glucose transporter 1 promotes the malignant phenotype of non-small cell lung cancer through integrin β1/Src/FAK signaling”. In: *Journal of Cancer* 10.20 (2019), p. 4989.

[183] Young Wha Koh, Su Jin Lee, and Seong Yong Park. “Differential expression and prognostic significance of GLUT1 according to histologic type of non-small-cell lung cancer and its association with volume-dependent parameters”. In: *Lung Cancer* 104 (2017), pp. 31–37.

[184] Min Yu et al. “The prognostic value of GLUT1 in cancers: a systematic review and meta-analysis”. In: *Oncotarget* 8.26 (2017), p. 43356.

[185] Qinlong Li et al. “Keratin 13 expression reprograms bone and brain metastases of human prostate cancer cells”. In: *Oncotarget* 7.51 (2016), p. 84645.

[186] Tam Quang Nguyen et al. “Enhanced KRT13 gene expression bestows radiation resistance in squamous cell carcinoma cells”. In: *In Vitro Cellular & Developmental Biology-Animal* (2021), pp. 1–15.

[187] Christian Rolfo and Luis Raez. “New targets bring hope in squamous cell lung cancer: neurotrophic tyrosine kinase gene fusions”. In: *Laboratory investigation* 97.11 (2017), pp. 1268–1270.

[188] Derek Wong, Stephen Yip, and Poul H Sorensen. “Methods for identifying patients with tropomyosin receptor kinase (TRK) fusion cancer”. In: *Pathology & Oncology Research* 26.3 (2020), pp. 1385–1399.

[189] Keigo Ozono et al. “Brain-derived neurotrophic factor/tropomyosin-related kinase B signaling pathway contributes to the aggressive behavior of lung squamous cell carcinoma”. In: *Laboratory Investigation* 97.11 (2017), pp. 1332–1342.

[190] Siyang Zhang et al. “TrkB is highly expressed in NSCLC and mediates BDNF-induced the activation of Pyk2 signaling and the invasion of A549 cells”. In: *BMC cancer* 10.1 (2010), pp. 1–8.

[191] Feng Zhang et al. “Identification of key transcription factors associated with lung squamous cell carcinoma”. In: *Medical science monitor: international medical journal of experimental and clinical research* 23 (2017), p. 172.

[192] Caiqi Cheng, Zhisen Ai, and Linyong Zhao. “Comprehensive analysis of the expression and prognosis for TFAP2 in human lung carcinoma”. In: *Genes & Genomics* 42 (2020), pp. 779–789.

[193] Jun Lu et al. “Chromatin accessibility analysis reveals that TFAP2A promotes angiogenesis in acquired resistance to anlotinib in lung cancer cells”. In: *Acta pharmacologica Sinica* 41.10 (2020), pp. 1357–1365.

[194] Ya-Ling Hsu et al. “Identification of novel gene expression signature in lung adenocarcinoma by using next-generation sequencing data and bioinformatics analysis”. In: *Oncotarget* 8.62 (2017), p. 104831.

[195] De Zeng, Haoyu Lin, Jianxiong Cui, and Weiquan Liang. “TOX3 is a favorable prognostic indicator and potential immunomodulatory factor in lung adenocarcinoma”. In: *Oncology letters* 18.4 (2019), pp. 4144–4152.

[196] Francisco Carrillo-Perez et al. “Machine-Learning-Based Late Fusion on Multi-Omics and Multi-Scale Data for Non-Small-Cell Lung Cancer Diagnosis”. In: *Journal of Personalized Medicine* 12.4 (2022), p. 601.

[197] Zhe-Wei Qiu, Jia-Hao Bi, Adi F Gazdar, and Kai Song. “Genome-wide copy number variation pattern analysis and a classification signature for non-small cell lung cancer”. In: *Genes, Chromosomes and Cancer* 56.7 (2017), pp. 559–569.

[198] Zheng Ye, Bo Sun, and Zhongdang Xiao. “Machine learning identifies 10 feature miRNAs for lung squamous cell carcinoma”. In: *Gene* 749 (2020), p. 144669.

[199] Zhihua Cai et al. “Classification of lung cancer using ensemble-based feature selection and machine learning methods”. In: *Molecular BioSystems* 11.3 (2015), pp. 791–800.

[200] Yan-ying Wang, Tao Ren, Ying-yun Cai, and Xiao-ye He. “MicroRNA let-7a inhibits the proliferation and invasion of nonsmall cell lung cancer cell line 95D by regulating K-Ras and HMGA2 gene expression”. In: *Cancer Biotherapy and Radiopharmaceuticals* 28.2 (2013), pp. 131–137.

[201] Ji-guang Zhang et al. “MicroRNA-21 (miR-21) represses tumor suppressor PTEN and promotes growth and invasion in non-small cell lung cancer (NSCLC)”. In: *Clinica chimica acta* 411.11-12 (2010), pp. 846–852.

[202] Hong Zheng, Alexandre Momeni, Pierre-Louis Cedoz, Hannes Vogel, and Olivier Gevaert. “Whole slide images reflect DNA methylation patterns of human tumors”. In: *NPJ genomic medicine* 5.1 (2020), pp. 1–10.

[203] Daniel Castillo-Secilla et al. “KnowSeq R-Bioc package: The automatic smart gene expression tool for retrieving relevant biological knowledge”. In: *Computers in Biology and Medicine* 133 (2021), p. 104387.

[204] Zhiyu Yang, Hongkun Yin, Lei Shi, and Xiaohua Qian. “A novel microRNA signature for pathological grading in lung adenocarcinoma based on TCGA and GEO data”. In: *International journal of molecular medicine* 45.5 (2020), pp. 1397–1408.

[205] Nan Shen et al. “A diagnostic panel of DNA methylation biomarkers for lung adenocarcinoma”. In: *Frontiers in oncology* 9 (2019), p. 1281.

[206] Olivier Gevaert, Robert Tibshirani, and Sylvia K Plevritis. “Pancancer analysis of DNA methylation-driven genes using MethylMix”. In: *Genome biology* 16.1 (2015), pp. 1–13.

[207] Tzong-Yi Lee, Kai-Yao Huang, Cheng-Hsiang Chuang, Cheng-Yang Lee, and Tzu-Hao Chang. “Incorporating deep learning and multi-omics autoencoding for analysis of lung adenocarcinoma prognostication”. In: *Computational Biology and Chemistry* 87 (2020), p. 107277.

[208] Francisco Carrillo-Perez et al. “Non-small-cell lung cancer classification via RNA-Seq and histology imaging probability fusion”. In: *BMC bioinformatics* 22.1 (2021), pp. 1–19.

[209] Ana Conesa et al. “A survey of best practices for RNA-seq data analysis”. In: *Genome biology* 17.1 (2016), pp. 1–19.

[210] Pauli Virtanen et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[211] Robi Polikar et al. “An ensemble based data fusion approach for early diagnosis of Alzheimer’s disease”. In: *Information Fusion* 9.1 (2008). Special Issue on Applications of Ensemble Methods, pp. 83–95.

[212] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[213] Benoıˆet Schmauch et al. “A deep learning model to predict RNA-Seq expression of tumours from whole slide images”. In: *Nature communications* 11.1 (2020), pp. 1–15.

[214] Inke R Koenig, Oliver Fuchs, Gesine Hansen, Erika von Mutius, and Matthias V Kopp. “What is precision medicine?” In: *European respiratory journal* 50.4 (2017).

[215] Djihad Hadjadj, Shriya Deshmukh, and Nada Jabado. “Entering the era of precision medicine in pediatric oncology”. In: *Nature Medicine* 26.11 (2020), pp. 1684–1685.

[216] Hidewaki Nakagawa and Masashi Fujita. “Whole genome sequencing analysis for cancer genomics and precision medicine”. In: *Cancer science* 109.3 (2018), pp. 513–522.

[217] Graham R Bignell et al. “Signatures of mutation and selection in the cancer genome”. In: *Nature* 463.7283 (2010), pp. 893–898.

[218] Lucas Schneider et al. “Integration of deep learning-based image analysis and genomic data in cancer pathology: A systematic review”. In: *European Journal of Cancer* 160 (2022), pp. 80–91.

[219] Da Zhang and Mansur Kabuka. “Multimodal deep representation learning for protein interaction identification and protein family classification”. In: *BMC bioinformatics* 20.16 (2019), pp. 1–14.

[220] Richard J Chen et al. “Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis”. In: *IEEE Transactions on Medical Imaging* (2020).

[221] Tom Brown et al. “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[222] Antonia Creswell et al. “Generative adversarial networks: An overview”. In: *IEEE Signal Processing Magazine* 35.1 (2018), pp. 53–65.

[223] Ruoqi Wei and Ausif Mahmood. “Recent advances in variational autoencoders with representation learning for biomedical informatics: A survey”. In: *Ieee Access* 9 (2020), pp. 4939–4956.

[224] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. "Hierarchical text-conditional image generation with clip latents". In: *arXiv preprint arXiv:2204.06125* (2022).

[225] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[226] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 6840–6851.

[227] Aditya Ramesh et al. "Zero-shot text-to-image generation". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 8821–8831.

[228] Ming Tao et al. "DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 16515–16525.

[229] Chitwan Saharia et al. "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding". In: *arXiv preprint arXiv:2205.11487* (2022).

[230] Jean-Baptiste Alayrac et al. "Flamingo: a Visual Language Model for Few-Shot Learning". In: *arXiv preprint arXiv:2204.14198* (2022).

[231] Tanya Barrett et al. "NCBI GEO: archive for functional genomics data sets—update". In: *Nucleic acids research* 41.D1 (2012), pp. D991–D995.

[232] Charlotte N Jennings et al. "Bridging the gap with the UK Genomics Pathology Imaging Collection". In: *Nature Medicine* (2022).

[233] Jonathan Ho et al. "Cascaded diffusion models for high fidelity image generation". In: *Journal of Machine Learning Research* 23.47 (2022), pp. 1–33.

[234] Yeping Lina Qiu, Hong Zheng, and Olivier Gevaert. "Genomic data imputation with variational auto-encoders". In: *GigaScience* 9.8 (2020), giaa082.

[235] Gregory P Way and Casey S Greene. "Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders". In: *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2018: Proceedings of the Pacific Symposium*. World Scientific. 2018, pp. 80–91.

[236] Ramon Viñas, Helena Andrés-Terré, Pietro Liò, and Kevin Bryson. "Adversarial generation of gene expression data". In: *Bioinformatics* 38.3 (2022), pp. 730–737.

[237] Adalberto Claudio Quiros et al. "Adversarial Learning of Cancer Tissue Representations". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 602–612.

[238] Maria Suntsova et al. "Atlas of RNA sequencing profiles for normal human tissues". In: *Scientific data* 6.1 (2019), pp. 1–9.

[239] Ming Y Lu et al. "AI-based pathology predicts origins for cancers of unknown primary". In: *Nature* 594.7861 (2021), pp. 106–110.

[240] Ming Y Lu et al. "Data-efficient and weakly supervised computational pathology on whole-slide images". In: *Nature biomedical engineering* 5.6 (2021), pp. 555–570.

[241] Nobuyuki Otsu. "A threshold selection method from gray-level histograms". In: *IEEE transactions on systems, man, and cybernetics* 9.1 (1979), pp. 62–66.

[242] Avik Pal and Aniket Das. "TorchGAN: A Flexible Framework for GAN Training and Evaluation". In: *Journal of Open Source Software* 6.66 (2021), p. 2606.

[243] Leland McInnes, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426* (2018).

[244] Katja Schwarz, Yiyi Liao, and Andreas Geiger. "On the frequency bias of generative models". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 18126–18136.

[245] Maria Angeles Peinado. "Histology and histochemistry of the aging cerebral cortex: an overview". In: *Microscopy research and technique* 43.1 (1998), pp. 1–7.

[246] Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. "Unrolled generative adversarial networks". In: *arXiv preprint arXiv:1611.02163* (2016).

[247] Tim Salimans et al. "Improved techniques for training gans". In: *Advances in neural information processing systems* 29 (2016).

[248] John Jumper et al. "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873 (2021), pp. 583–589.

[249] Darren J Burgess. "Spatial transcriptomics coming of age". In: *Nature Reviews Genetics* 20.6 (2019), pp. 317–317.

[250] Anjali Rao, Dalia Barkley, Gustavo S França, and Itai Yanai. "Exploring tissue architecture using spatial transcriptomics". In: *Nature* 596.7871 (2021), pp. 211–220.

[251] Yuansong Zeng et al. "Spatial transcriptomics prediction from histology jointly through Transformer and graph neural networks". In: *Briefings in Bioinformatics* 23.5 (2022), bbac297.

[252] Yanning Zhou et al. "Cgc-net: Cell graph convolutional network for grading of colorectal cancer histology images". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 2019.

[253] Yongju Lee et al. "Derivation of prognostic contextual histopathological features from whole-slide images of tumours via graph deep learning". In: *Nature Biomedical Engineering* (2022), pp. 1–15.

[254] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[255] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. "Do vision transformers see like convolutional neural networks?" In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 12116–12128.

[256] Narmin Ghaffari Laleh et al. "Adversarial attacks and adversarial robustness in computational pathology". In: *Nature communications* 13.1 (2022), pp. 1–10.

[257] Richard J Chen et al. "Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 16144–16155.

[258] Adalberto Claudio Quiros et al. “Self-supervised learning unveils morphological clusters behind lung cancer types and prognosis”. In: *arXiv preprint arXiv:2205.01931* (2022).