UNIVERSIDAD NACIONAL DE COLOMBIA

A Data-driven Representation Learning for Tumor Tissue Differentiation from Non-Small Cell Lung Cancer Histopathology Images

Fabian Alberto Cano Ramirez

Universidad Nacional de Colombia
Facultad de Medicina, Departamento de Imágenes Diagnósticas
Bogotá D.C., Colombia
2022

A Data-driven Representation Learning for Tumor Tissue Differentiation from Non-Small Cell Lung Cancer Histopathology Images

Fabian Alberto Cano Ramirez

Thesis presented as a partial requirement to qualify for the title of:
Master in Biomedical Engineering

Advisor:
Eduardo Romero Castro, M.Sc., Ph.D.
Co-advisor:
Angel Alfonso Cruz Roa, M.Sc., Ph.D.

Research field:
Biomedical Engineering
Research group:
Computer Imaging and Medical Applications Laboratory - CIM@LAB

Universidad Nacional de Colombia
Facultad de Medicina, Departamento de Imágenes Diagnósticas
Bogotá D.C., Colombia
2022

A mis padres y a mi hermana
por todo el apoyo y motivación que me han dado
en cada paso de mi vida
Sin ellos, nada de esto habría sido posible

"Si tienes miedo a intentarlo, es porque lo estás deseando. El miedo te indica, claramente, que debes hacerlo".

— **Luis Ramiro**

Acknowledgments

I want to express my gratitude to my advisor, Professor Eduardo Romero, for his guidance and patience, for all the support, and for the many academic and personal lessons that he has taught me in all this time. I have learned many skills and abilities thanks to him, always motivating me not to give up and to go one step beyond the limits. My thanks also goes to Professor Angel Cruz-Roa, because his vision, support and advice have accompanied my training process, in addition to the achievements and products I have made. From both I have learned that it is not important to know everything, what matters is the attitude to continue learning and share that learning with the people who need it.

I also want to thank Charlems Alvarez who welcomed me from the beginning and was attentive to each of my steps, providing ideas, advice and company to solve the problems that arose. Thanks to all the colleagues and friends of the Cim@Lab research group. My most sincere thanks to Charlems Alvarez, Jennifer Salguero, Mauricio Caviedes, Eileen Montoya, Maria Jaramillo, and Andres Sandino. Also thanks to the Universidad Nacional de Colombia, its professors and infrastructure. To the Master's program in Biomedical Engineering. To the Universidad de los Llanos, and to the AdaLab research hotbed, for the spaces to develop new knowledge and carry out ideas.

Finally, a very special thanks to my parents, who supported me to carry out my studies, always with their company and motivation. Thanks also to my sister, who was always there to give me peace of mind and share moments of joy. Many sacrifices were made to meet this goal, without them none of this would have been possible.

This thesis has been funded by the project BPIN 2019000100060 "Implementación de una Red de Investigación, Desarrollo Tecnológico e Innovación en Patología Digital (RedPat) soportada por tecnologías de la Industria 4.0" of FCTeI of SGR resources, which was approved by OCAD of FCTeI and MinCiencias .

Abstract

A Data-driven Representation Learning for Tumor Tissue Differentiation from Non-Small Cell Lung Cancer Histopathology Images

Lung cancer is the second most common type and the leading cause of cancer death in the world. It is divided into different types according to cellular and tissular features, and in turn, these types are distinguished by typical patterns that represent them. Each histological subtype of lung cancer is associated with the prognosis and treatment of patients, and is subjectively stratified mainly by its morphological features. However, due to the very nature of the disease, this stratification varies since there is no specialized grading system, and also because of the difficulty of characterizing cases that generally contain mixtures of histological patterns and unspecified tissues, which therefore, alters the diagnosis and prognosis of patients. This research work addresses a computational data-driven strategy to characterize histological patterns of lung cancer, in addition to determining its differentiation and aggressiveness, in order to support decision-making in clinical practice. Therefore, this work has been divided in two parts. The first part presents a supervised subtype differentiation learning of lung cancer features in a latent space constructed with a variational autoencoder. In such space, complicated patterns are quantified by estimating a differentiation grade of typical encoded features of lung cancer subtypes. Then, a logistic regression model assigns differentiation cancer subtype grade to the embedded tissue samples. This approach builds up a subtype differentiation grade of non-small cell lung cancer among complex structures which are fully interpretable and integrable with a pathology workflow. Finally, the second part presents an unsupervised computational approach based on an ensemble of tissue-specialized variational autoencoders, which were trained per histopathology subtype, to build an unsupervised embedded tissue-image representation. This representation was used to train a Random Forest classifier of three lung adenocarcinoma histology subtypes (lepidic, papillary and solid), and a 2D-visually interpretable projection from the learned embedded representation.

Keywords: Digital Pathology, Tissue Representation, Histopathology, Variational Autoencoder, Lung Adenocarcinoma, Lung Cancer

Resumen

Un aprendizaje de representación basado en datos para la diferenciación de tejido tumoral a partir de imágenes de histopatología de cáncer de pulmón de células no pequeñas

El cáncer de pulmón es el segundo tipo más común y la principal causa de muerte por cáncer en el mundo. Se divide en diferentes tipos según las características celulares y tisulares, y a su vez, estos tipos se distinguen por los patrones histológicos típicos que los representan. Cada subtipo histológico de cáncer de pulmón se asocia con el pronóstico y tratamiento de los pacientes, y se estratifica subjetivamente por parte de los patólogos principalmente por sus características morfológicas. Sin embargo, por la propia naturaleza de la enfermedad, esta estratificación varía ya que no existe un sistema de gradación especializado, y también por la dificultad de caracterizar los casos que generalmente contienen mezclas de patrones histológicos y tejidos no especificados, lo que puede afectar la precisión del diagnóstico y pronóstico de los pacientes. Este trabajo de investigación aborda una estrategia computacional basada en datos para caracterizar los patrones histológicos del cáncer de pulmón, además de determinar su diferenciación y agresividad, con el fin de apoyar la toma de decisiones en la práctica clínica. Por ello, este trabajo se ha dividido en dos partes. La primera parte presenta un aprendizaje supervisado de diferenciación de subtipos de características de cáncer de pulmón en un espacio latente construido con un autocodificador variacional. En dicho espacio, los patrones complejos se cuantifican mediante la estimación de un grado de diferenciación de las características codificadas típicas de los subtipos de cáncer de pulmón. Luego, un modelo de regresión logística asigna un grado de diferenciación del subtipo de cáncer a las muestras de tejido codificadas. Este enfoque construye un grado de diferenciación de subtipos de cáncer de pulmón de células no pequeñas entre estructuras complejas que son totalmente interpretables e integrables con un flujo de trabajo de patología. Finalmente, la segunda parte presenta un enfoque computacional no supervisado basado en un conjunto de codificadores automáticos variacionales especializados en tejidos, que fueron entrenados por subtipo de histopatología, para construir una representación de imagen de tejido codificada no supervisada. Esta representación se usó para entrenar un clasificador Random Forest para distinguir entre tres subtipos histológicos de adenocarcinoma de pulmón (lepídico, papilar y sólido) y una proyección visualmente interpretable en 2D a partir de la representación incrustada aprendida.

Palabras clave: Patología Digital, Representación de tejidos, Histopatología, Autocodificador Variacional, Adenocarcinoma de pulmón, Cáncer de pulmón.

Contents

List of Figures

List of Tables

1 Introduction

According to the most recent report from the Global Cancer Observatory (GLOBOCAN), lung cancer was the second most common type of cancer in 2020, accounting for 11.4% of all cancer cases, but was the leading cause of cancer deaths with 18% of all deaths worldwide, likewise, the 5-year survival of patients with lung cancer is estimated to be 10 to 20 percent for most countries [8]. Lung cancer is mainly divided into two types: *non-small cell lung cancer* (NSCLC), causing about of 85% of all diagnosed cases (see Figure **1-1**), and *small cell lung cancer* (SCLC), with an incidence about of 15% [8, 9]. SCLC is the most aggressive form of lung cancer and is closely related to smoking. These types of tumors have a worse prognosis because the cancer cells quickly spread to other parts of the body. On the other hand, NSCLC is more frequent in non-smokers and is associated with a better prognosis [1].

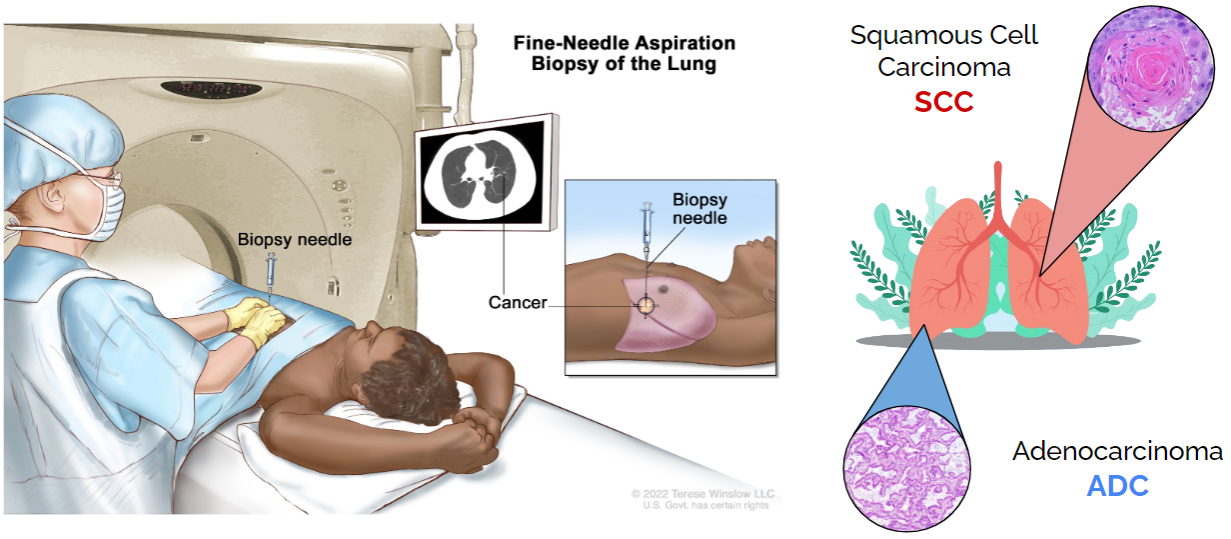

Figure 1-1: *Left side:* Fine-needle aspiration biopsy of the lung. A small piece of tissue is removed using several computed tomography (CT) images and is then examined under a microscope. *Right side:* The main types of NSCLC and their common location. Adenocarcinoma is located mainly in the peripheral zone and in the upper lobes of the lung. Squamous cell carcinoma is located mainly in central areas, originating in the bronchi [1]. Adapted from: Lung Biopsy. National Cancer Institute. Accessed November 25, 2022.

However, only a small proportion of patients with NSCLC, less than 20%, are diagnosed in early stages of the disease whereas in most cases, more than 47%, are diagnosed in later stages (stage III/IV), where lymph nodes or distant organs are involved, significantly affecting their survival [10].

1.1 Lung Cancer differentiation

NSCLC is subdivided into several histological subtypes (see Figure **1-2**), with *adenocarcinoma* (ADC) being one of the most frequently diagnosed among them, whose incidence is estimated to be close to 40%; followed by *squamous cell carcinoma* (SCC), with an incidence around 30% of all cases [11, 12, 13]. Several studies have demonstrated the importance of differentiating NSCLC subtypes [14] since available treatments are different depending on the histological subtype [15]. Also, early characterization of NSCLC is crucial to determine patient prognosis and survival [16].

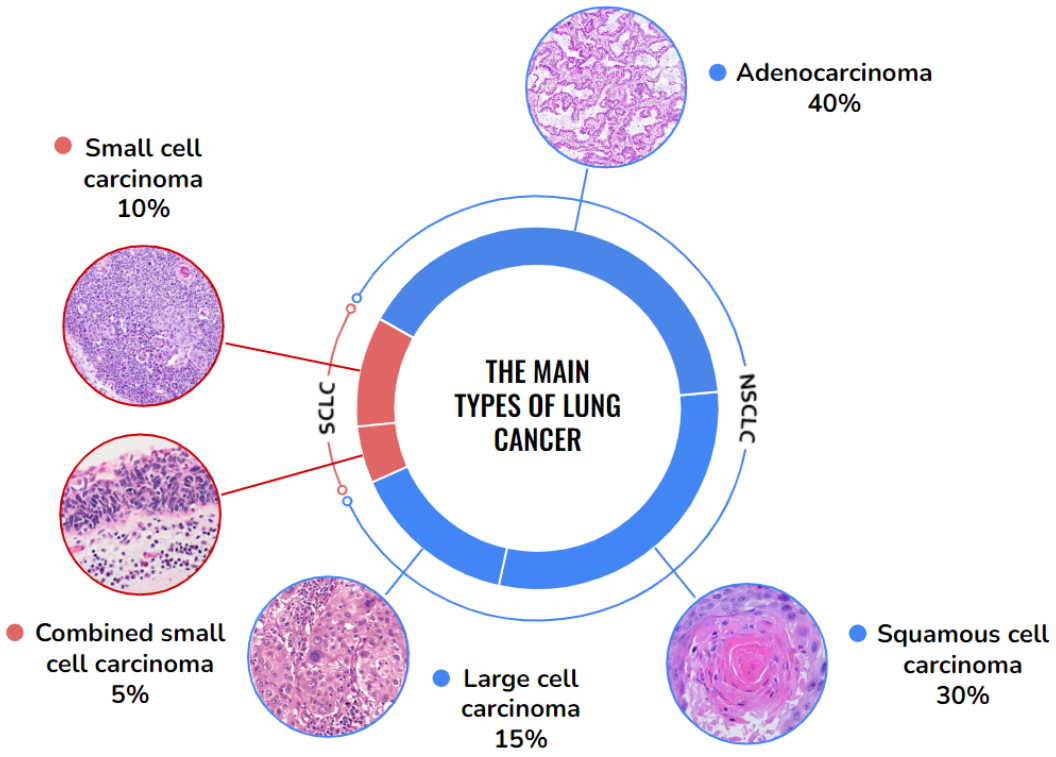

Figure 1-2: The main types of lung cancer and their estimated incidence. Adapted from: Bender, E. (2014) [2]

Patient prognosis is estimated according to the tumor grade and is typically based on its architecture and cellular features. In other types of cancer, there are well-established tumor grading methods associated with prognosis, such as the case of the Gleason Score for prostate cancer [17]. However, this is not the case in lung cancer because grading is subjective by

pathologists and it is currently determined on a spectrum from well-differentiated (grade 1) to the most poorly-differentiated (grades 3 or 4), also, grading lacks any specific guideline or assessment system that allows its differentiation [18]. For this reason, recent studies have explored the relationship between histological subtypes of NSCLC and its aggressiveness [14, 19, 20]. However, the design of a quantitative measure that reflects the differentiation between histological subtypes of NSCLC and the need to characterize them remains a challenge.

Pathologists analyze tissue samples under the microscope by looking for cellular and tissular patterns and morphology to describe and interpret the findings in order to determine the diagnosis and tumor subtypes [21]. Thus, the histological subtype of NSCLC and its infiltration are generally described in the cancer diagnosis. Because of this, it has been reported that ADC, the main diagnosed subtype of NSCLC, currently is mainly divided into five histological subtypes (see Figure **1-3**), which include *lepidic, acinar, papillary, micropapillary and solid* [22]. It has been observed that histological subtypes of ADC are related to prognosis, reason why the World Health Organization (WHO) recommends classifying ADC into four grades: *well-differentiated* (Grade 1), *moderately-differentiated* (Grade 2), and *poorly-differentiated* (Grades 3 or 4), based on histological subtype and its infiltration, as shown in the Table **1-1**. However, subtype characterization and tissue grading are intrinsically subjective and vary due to a lack of a specialized grading system and pathologist expertise [15, 23, 24].

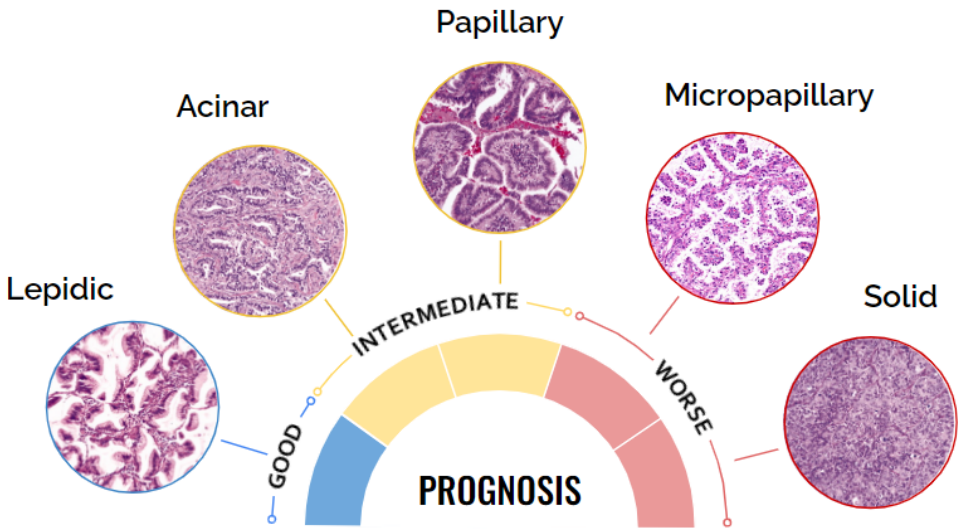

Figure 1-3: The five main histological subtypes of ADC correlate with its most common prognosis. The lepidic pattern is the least aggressive and has the best prognosis, the acinar and papillary patterns are considered intermediate, and the micropapillary and solid patterns are the most aggressive and are generally associated with the worst prognosis. Adapted from: Kuhn, E. (2018) [3].

Table 1-1: Histopathologic grading scheme for ADC proposed by the International Association for the Study of Lung Cancer Pathology Committee (IASLC) [6]. However, since grading systems have not been established for all histological types of lung cancer, its reproducibility and prognostic significance have not been rigorously tested [6, 7].

Grade	Description
Grade X[a]	Cannot be assessed
Grade 1	Well-differentiated (lepidic-predominant with no or $< 20\%$ high-grade pattern)
Grade 2	Moderately differentiated (acinar or papillary-predominant with no or $< 20\%$ high-grade pattern)
Grade 3	Poorly differentiated (any tumor with $\geq 20\%$ high-grade pattern (i.e. solid, micropapillary, cribriform, or complex glandular pattern)
Grade 4[a]	Undifferentiated

[a] Grading scheme used for the American Joint Committee on Cancer (AJCC) [7].

Additionally, the accurate interpretation can occasionally be a challenge, particularly in poorly-differentiated cases [25], in which cells preserve a normal appearance, and generally tend to grow slowly. In contrast, in high-grade cancers, cells grow rapidly and abnormally in different regions, resulting in a worse prognosis, and may require different treatments depending on the subtype and tissue infiltration [26]. The incidence of NSCLC varies widely, and it is common that there are not enough samples for each histological subtype. This is due to the inherent difficulty of characterizing some tissue samples as they usually contain mixtures of histological patterns and unspecified tissues [27]. In fact, the mixture of histological subtypes is challenging and alters the patient diagnosis and prognosis [28].

1.1.1 Morphological features

In lung cancer, radiological examinations are the first line to determine the nature of the disease, since they are non-invasive methods. However, the main disadvantage of this type of imaging is that it does not provide enough information to accurately determine the extent and characteristics of the disease, so histopathological analysis remains the gold standard in the diagnosis of cancer [15].

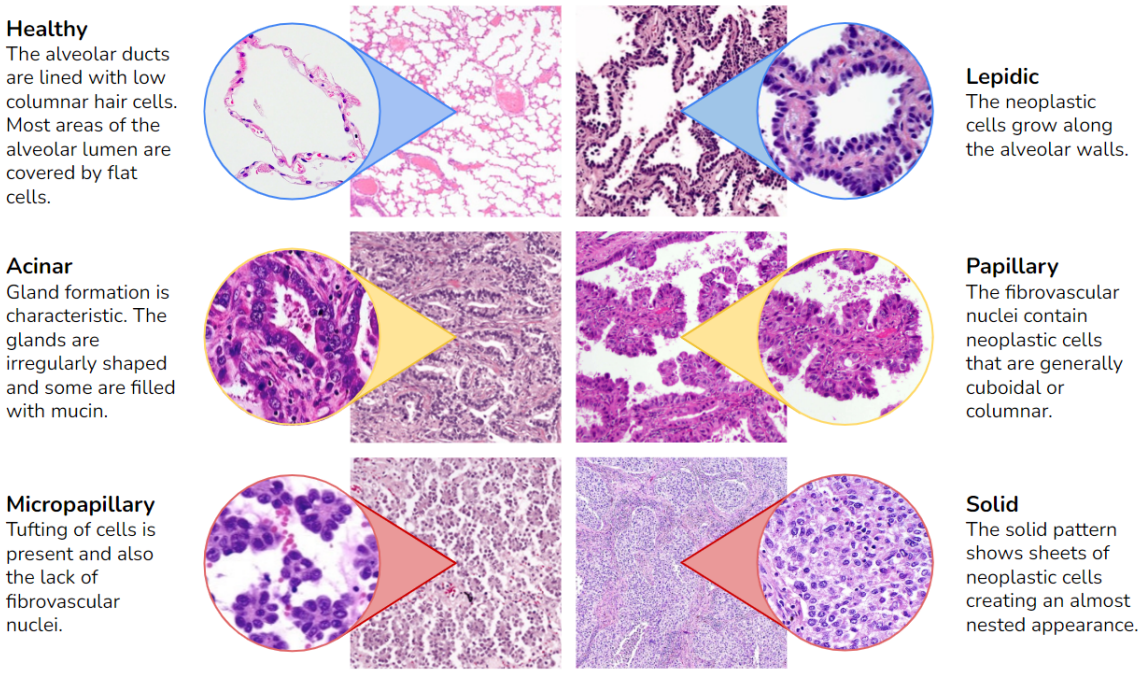

Figure 1-4: Comparison of the main morphological features of healthy lung tissue and the histological subtypes of ADC. Adapted from: *Adenocarcinoma overview*, by Pathology Outlines, 2019 [4].

Tumor tissue grade and its aggressiveness are determined by cellular and tissular features (see Figure **1-4**) [21]. The *lepidic* pattern is composed of generally soft cells that grow along the alveolar walls, and in general, the invasive foci contain characteristics such as malignant glands, too-small alveoli, angulated or branched, and the presence of an interstitial desmoplastic reaction. The *acinar* pattern is mainly composed of neoplastic glands arranged in acini. These structures can be small tubules, angulated and branching cords, and even more complex irregular glands. Occasionally, the acini collapse so that the acinar pattern is difficult to recognize. The *papillary* pattern contains neoplastic cells lining fibrovascular nuclei of variable size and branching, whose neoplastic cells are generally cuboidal or columnar. The *micropapillary* pattern is mainly composed of papillary tufts that can fold on the alveolar surface, float within the alveoli, and even infiltrate the stroma. The *solid* pattern is composed of solid nests, sometimes with a vaguely scaly appearance. The cell cytoplasm can be light, dark, eosinophilic, or basophilic, and the nuclei are usually highly pleomorphic. In some cases, immunohistochemistry must be applied to differentiate it from other subtypes such as non-keratinizing SCC [3, 29].

1.1.2 Incidence and predominance

Currently, there are no detailed global studies about the incidence of histological subtypes of ADC and its differentiation, hence these incidences can only be estimated in isolated

studies. For example, in Gengpeng *et al.* [30], lepidic and acinar predominant subtypes are analyzed in a cohort about of two thousand patients, where the incidence was 0.72% and 0.45% respectively. In Kadota *et al.* [31], a retrospective study is performed with about one thousand patients diagnosed with ADC and classified according to the predominance of their histological subtypes, finding that 10% of the cases had lepidic predominance, 40% acinar, 23% papillary, 6% micropapillary, 13% solid and 4% mucinous. In Kinno *et al.* [32], about of two thousand patients were selected in which 34.8% were characterized with predominant papillary histological subtype, 19.7% predominant lepidic, 12.7% predominant solid, 9.3% predominant acinar, 6.3% predominant micropapillary, and 3.9% mucinous. Characterization of histological subtypes of ADC is critical to guide patient treatment and prognosis. However, this task is often challenging due to the heterogeneous nature of ADC and the subjective criteria for evaluation, added to the proportion of difficult cases, such as those mixed, that is, those that present two or more predominant patterns, and the cases with unspecified or poorly differentiated patterns.

Due to the heterogeneity of NSCLC data, a collaboration between pathologists and engineers is typically required to build artificial intelligence (AI) models that exploit the potential of the data. *Data-driven models* have the advantage of offering a greater interpretation capacity thanks to the knowledge provided by pathologists, a feature that differentiates them from the main deep learning (DL) models. While the latter models are still considered a black box in understanding the nature of the disease, data-driven models allow the active participation of experts to, among other things, define the regions of interest and interpret the results [15].

1.2 Deep Learning in lung cancer

Computational approaches have demonstrated the potential to learn the typical patterns that characterize the histological subtypes of NSCLC, as well as, to guide pathologists and oncologists for improving the accuracy of medical diagnosis [33]. The development of computational pathology tools for quantitative diagnosis support and histological tumor subtypes characterization from histopathology digital images could help to reduce the time needed to identify and interpret findings, allowing pathologists to spend time in other aspects of clinical and pathological workflow or cancer research, such as image-based biomarker interpretation and discovery, or design a tissue-based grading system to improve the estimation of patient prognosis [34, 35, 36].

The current rise of digital pathology and digitized histopathology slides have made possible to integrate different AI and machine learning (ML) models to analyze and perform tasks for diagnosis support. Several works have used computational models in pathology with data from the most common types of cancer such as breast [37, 38] or prostate [39, 40]. In most

of them, convolutional neural networks (CNN) have been implemented because their ability to learn all kinds of patterns on images. For example, Litjens *et al.* [41], trained a CNN with prostate and breast cancer cases to improve histological diagnosis in biopsies with early metastases. Veeling *et al.* [42] proposed a CNN model to detect tumors in a set of lymph node metastasis images; and Mukhopadhyay *et al.* [43], selected a sample of patients and developed different automatic methods to perform the main diagnostic tasks, concluding that the results obtained by the models were above those obtained in traditional methods involving microscopic examination.

Computational models designed to perform different tasks, such as classification, detection or segmentation, using data from NSCLC subtypes have grown progressively [44]. The most common CNN-based models are pretrained architectures from natural images (e.g., ImageNet) such as ResNet50 or InceptionV3 tuned to perform tasks such as NSCLC histological subtype classification [14, 45]. Other more specialized ones, such as UNet [46], are used for tissue-type segmentation tasks in digital histopathology images [47]. Although there are many applications, these types of DL models are generally not interpretable [48], so their acceptance in clinical practice is still a matter of discussion. In addition, many of them depend on a large amount of annotated data to be trained by supervised approaches and get a good performance [49]. However, in some applications, such as cancer histopathology images, large amounts of annotated data are not available, which is why some recent methods have made efforts to make the most of the small amount of data available to learn its features and perform several tasks. These DL models are generally considered unsupervised [50, 51, 52], semi-supervised [53, 54] or data-driven models [55, 56, 57], since they take full advantage of the complexity and representativeness of the small amount of data, obtaining results comparable to classical supervised methods.

1.2.1 Histological subtype classification

Particularly, in lung cancer, the characterization of histological subtypes of NSCLC has been explored. In Coudray *et al.* [14], a deep learning model for automatic analysis of histological subtypes of NSCLC was developed with data retrieved from The Cancer Genome Atlas (TCGA) and own cohorts. The data include images that correspond to frozen tissues and associated genetic information. The authors demonstrate the capacity of a CNN model to support lung cancer diagnosis in difficult-to-diagnose cases, and also the associated genetic information contributes to improve the model performance by providing additional features to characterize the disease.

In other approaches, Gertych *et al.* [19] implemented a CNN to characterize four histological patterns of ADC (acinar, micropapillary, solid, and cribriform). There, they compared the results obtained by the model with the interpretation described by pathologists and determined that machine learning models, such as CNN, have the potential to distinguish

the histological subtypes of ADC and their growth patterns, as well as to help pathologists to quantify them. Wei *et al.* [20] implemented a CNN to classify the main histological subtypes of ADC (lepidic, papillary, micro-papillary and solid) and compare the results obtained with the annotations made by pathologists. They concluded that their model had the potential to classify histological subtypes of ADC and assist pathologists in clinical practice by pre-scanning the image and highlighting regions of diagnostic interest. Yu *et al.* [58] developed a CNN for classifying NSCLC images using genetic information. Their main goal was to quantitatively characterize histopathology images to identify tumor regions and their histological subtype. Their approach objectively estimates histological patterns that characterize NSCLC, which in turn serves as a support system for pathologists in the evaluation of lung cancer.

1.2.2 Tissue segmentation

Given the high variability of histological subtypes of NSCLC, recent efforts have emerged to properly characterize them. Recently, international challenges have emerged in retrieving difficult-to-characterized NSCLC data available on TCGA. One of the most recent is the WSSS4LUAD [59], organized by The 5th International Symposium on Image Computing and Digital Medicine (ISICDM 2021), in which it was sought to characterize some types of tissue in ADC images, such as, tumor epithelial tissue, tumor-associated stroma tissue and normal tissue, using only image-level annotations with a set of 67 ADC images retrieved from Guangdong Provincial People Hospital (GDPH) and 20 ADC images retrieved from TCGA, one image per patient. Participants were expected to be able to design robust computational models using the small amount of available data to segment the three tissue types with the provided images. The designed computational models proved to be a feasible solution to replace the manual characterization of the tissues, being of great interest to reduce annotation efforts. In addition to the multiple applications and promising results in the characterization and differentiation of cancer, computational models in digital pathology are including pathologists and oncologists as guarantors of their operation and as support to develop more specialized models that allow understanding of the nature of pathologies [59, 20].

Computational models of DL contribute to improve the clinical workflow, since they allow the automation of different tasks that are normally time consuming, in addition to supporting decision-making of diagnostic interest for diseases such as NSCLC. Although the difficulties in applying this type of techniques in clinical practice are well known, regulations have recently been approved that allow its implementation [60], hence there is a growing interest in interpreting its results [61]. For this reason, unlike DL-based AI models, data-driven models combine visual representations of the data with the expertise of pathologists to interpret the results and thereby, among other things, stratify risk and predict response to

treatments [20, 62].

1.2.3 Unsupervised feature/image representation

Histopathology image analysis using AI models, particularly in DL field, has encouraged the growth of image analysis techniques that automatically extract relevant features using data-driven approaches [20]. The introduction of these types of techniques, and their ability to identify pathological features that contribute to diagnosis, prognosis, and prediction, could help pathologists to guide the patient treatment and prognosis [33]. The prognosis consists of stratifying patients taking into account the diagnosis or progression of the disease, by allowing clinicians and oncologists to make the necessary decisions for their correct treatment. In this sense, the potential of AI models lies in providing the necessary tools to identify complex patterns and support clinical decision-making that guides the selection of the appropriate treatment for each type of disease [15, 58].

Autoencoders

Autoencoders (AE) are a type of architecture used in dimensionality reduction processes (see Figure **1-5**). There are three main components of an AE: an encoder, a latent feature representation, and a decoder. In most typical architectures, the encoder and decoder are neural networks [63].

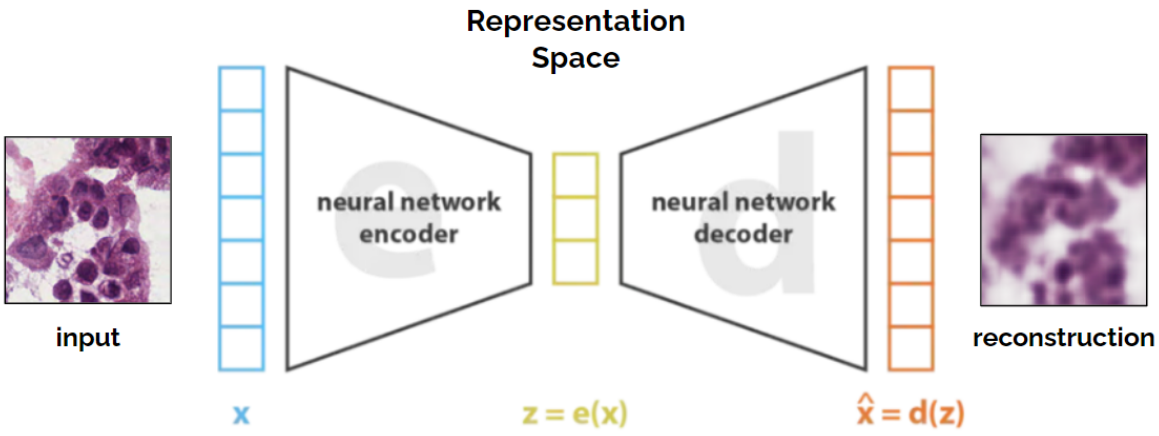

Figure 1-5: A typical architecture of an AE. The encoder e encodes the input information, from the original space to the representation space z, also called latent space, and then the decoder d decompresses the information. Adapted from Rocca (2019) [5].

The AE tries to reproduce the given input as an output. The learning procedure then consists of learning the encoder and decoder functions simultaneously, minimizing the reconstruction loss:

$$\text{Loss} = ||x - \hat{x}||^2 = ||x - \text{d}(z)||^2 = ||x - \text{d}(\text{e}(x))||^2 \tag{1-1}$$

where, the reconstruction loss is the difference between the model's output $\hat{x}$ and its input x, penalizing $d(e(x))$ for being dissimilar to x [64]. Depending on the input information, the compression can be lossy, that is, a part of the information is lost during the encoding process and cannot be recovered during the decoding process [5].

Variational Autoencoders

The Variational Autoencoder (VAE) is a variant of the classical AE architecture, where instead of a deterministic function, the latent space is regularized by encoding the input as a mixture of Gaussian distributions in order to return the mean and variance values that describe these Gaussians [65]. In this sense, the distributions returned by the encoder are enforced to be close to a known distribution, a standard normal distribution [64, 5].

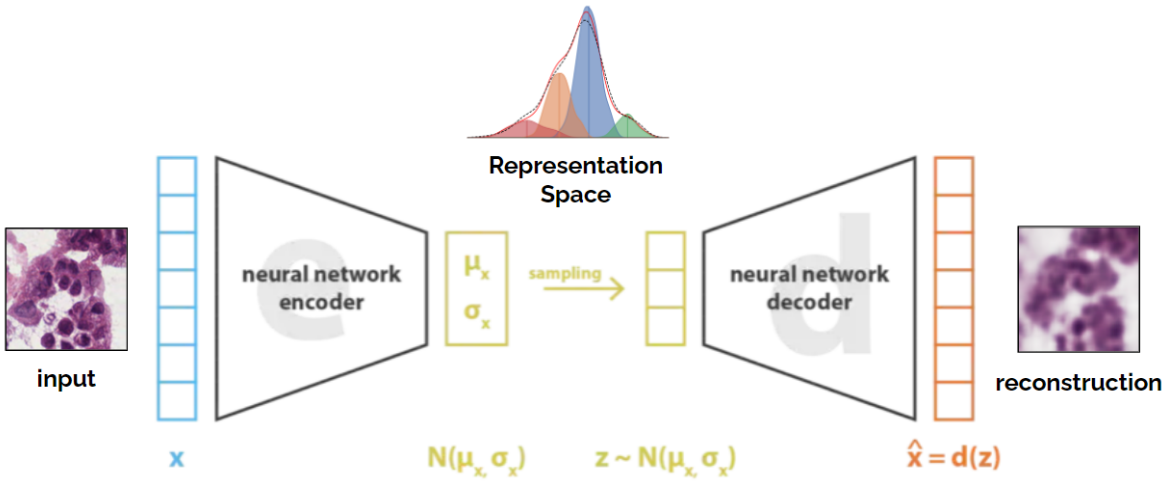

Figure 1-6: A typical architecture of a VAE. In this case, the latent space is regularized by encoding the input as a mixture of Gaussian distributions. Adapted from Rocca (2019) [5].

Therefore, e and d can also be thought of as conditional probability distributions. So, the loss function is composed of two terms: a reconstruction term and a regularization term. The reconstruction, as in an AE, is the difference between the output and the input of the model. Otherwise, the regularization term is the Kullback-Leibler (KL) divergence between the estimated probability distribution of the data and a reference distribution, a standard normal distribution. This function tries to force the encoder network to be as similar as possible to the reference distribution [64].

Loss function and regularization

The optimization objective of a VAE, as in other variational methods, is the evidence lower bound, abbreviated as ELBO. In general, the ELBO is derived through Jensen's inequality (see Annex A) [65]. It brings the distributions returned by the encoder close to a chosen distribution, e.g., normal standard:

$$\text{ELBO} = \underbrace{-\frac{1}{2}\sum_{j=1}^{J}(1+(\sigma_j^2)-(\mu_j^2)-(e^{\sigma_j^2}))}_{\text{Regularization}} + \underbrace{\sum_{i=1}^{D} x_i \log \hat{x}_i + (1-x_i)\log(1-\hat{x}_i)}_{\text{Reconstruction}} \tag{1-2}$$

where, the regularization term is the result of the approximation for a Gaussian assumption, and the reconstruction term is the definition of the binary cross entropy (see Annex A).

Additionally, if one wants to regularize the latent space taking into account the compromise between the reconstruction and the form of the regularized space, *Higgings et.al.* (2017) [66] explore an approach that introduces an adjustable hyperparameter β that modulates the learning constraints applied to the model, while balancing the trade-off between the model capability to reconstruct the input image and the understanding of representation space. If the value of the parameter $\beta = 1$, corresponds to the base model of the VAE [67], with $\beta > 1$, the model is pushed to learn a more efficient latent representation for the data, untangling the generated space. On the contrary, if $\beta < 1$ the model tries to reconstruct the input data in a better way.

1.3 Contributions and Academic Products

This work addresses a data-driven computational strategy to characterize the histological patterns of lung cancer, in addition to determining its differentiation and aggressiveness, in order to support decision-making in clinical practice. The main contributions of this work are the construction of a representation space based on the learning of a variational autoencoder in order to quantify the patterns that identify the subtypes of non-small cell lung cancer, in addition to estimating the differentiation grade of each subtype. Moreover, an ensemble of variational autoencoders specialized in several histological subtypes of lung adenocarcinoma is explored, which allows characterizing them from few learned features. The publications presented as part of this thesis are mentioned in the next section (see Section 1.3.1).

1.3.1 Academic products

Results of this work were published in:

- **Fabian Cano**, Charlems Alvarez-Jimenez, David Becerra, Andres Siabatto, Angel Cruz-Roa, Eduardo Romero. *A supervised subtype differentiation learning for building invariant features of non-small cell lung cancer in a latent space of a Variational Autoencoder*. Proc. SPIE 12088, 17th International Symposium on Medical Information Processing and Analysis - SIPAIM 2021, Campinas (Brazil), November 2021. doi.org/10.1117/12.2606255

- **Fabian Cano**, Charlems Alvarez-Jimenez, Eduardo Romero, Angel Cruz-Roa. *Ensemble of unsupervised learned image representations based on variational autoencoders for lung adenocarcinoma subtypes*. Submitted to the 20th IEEE International Symposium on Biomedical Imaging - ISBI 2023, Cartagena de Indias (Colombia), April 2023.

Additional related work:

- **Fabian Cano**, Angel Cruz-Roa. *Analysis of a semi-supervised learning algorithm of self-training based on convolutional neural networks in breast cancer histopathology images*. Submitted as poster to the 17th International Symposium on Medical Information Processing and Analysis - SIPAIM 2021, Campinas (Brazil), November 2021.

1.4 Organization of this thesis

The remaining chapters of this thesis are organized as follows:

- ***Chapter 2: A supervised subtype differentiation learning of NSCLC.*** This chapter presents a supervised subtype learning approach of histopathology tissue samples, using typical tissue samples, from ADC and SCC, selected by expert pathologists. The basic idea is to use a variational autoencoder to construct a latent space and to project onto it a selection of typical histopathological patterns which may describe objectively variations of these patterns. The ADC and SCC images of tissue samples were observed to map at different locations of the latent space and grouped around very different regions in a 2D projection. Finally, the differentiation grade is estimated for each example and stratified according to the typical features of each cancer subtype (ADC and SCC).

- ***Chapter 3: Ensemble of image representations for ADC subtypes.*** This chapter presents a computational pathology approach based on an ensemble of tissue-specialized variational autoencoders, each of them trained with the associated histopathology ADC subtype, i.e., lepidic, papillary and solid. This approach aims to build an unsupervised embedded tissue-image representation used to: 1) train a Random Forest tissue classifier of ADC subtypes, and 2) construct a 2D projection that is visually interpretable of tumor tissue sample distribution of ADC subtypes in the embedded

space. The proposed approach of unsupervised embedded tissue-image representation allows a good histopathology ADC subtypes differentiation, and a semantic spatial distribution of tissues into embedded space, placing typical isolated histopathology patterns of ADC subtypes (e.g. lepidic and solid) in the periphery, and untypical and mixtures of histopathology patterns (e.g. papillary) in the central zone. These two characteristics could help pathologists by providing more quantitative, objective and interpretable tools of computational pathology for diagnosis support in the more challenging and relevant tumor differentiation among histopathology ADC subtypes.

- ***Chapter 4: Conclusions and perspectives.*** This final chapter presents the main conclusions of this thesis, highlighting the main contributions, the most important findings and their impact on research and practical areas. Finally, future research directions and perspectives are presented and discussed.

2 A supervised subtype differentiation learning of NSCLC

This chapter presents a supervised subtype differentiation learning of lung cancer features in a latent space constructed with a variational autoencoder. Specifically, selected tissue samples of NSCLC are mapped to a latent space and a logistic regression model assigns differentiation cancer subtype grade to the embedded tissue samples. Typical tissue samples of well-differentiated lung cancer subtypes are grouped close in the latent space with high confidence of the differentiation grade, while poorly differentiated tissue samples, with lower confidence of the differentiation grade, are located at other latent space regions. The best variational autoencoder achieves an average performance of MAE = (0.072 ± 0.0004) and RMSE = (0.2654 ± 0.0019). These results demonstrate this type of representation may capture a reduced set of histopathological invariants, use them to quantify complex patterns and improve the reproducibility of certain estimations. A complete version of this chapter has been accepted for publication as a research article in the proceedings of **17th International Symposium on Medical Information Processing and Analysis** (see reference [68]).

2.1 Introduction

According to the World Health Organization (WHO), lung cancer was the most common type of cancer worldwide in 2018 [69]. It is mainly divided into two subtypes, being non-small cell lung cancer (NSCLC) the most frequent with about 85% of incidence [1]. This neoplasia is subdivided into adenocarcinoma (ADC) and squamous cell carcinoma (SCC) [11], each characterized by well-defined features. ADC is the most frequently diagnosed and is usually located at the lung periphery [12]. SCC is less frequent, with about 30% of incidence, and it is generally located at the central region of the lung, close to the bronchi, reason why it is considered more aggressive than ADC [12].

Several studies have shown how important is to differentiate these NSCLC subtypes [14], because the available treatments are different for each cancer subtype. In addition, early characterization of lung cancer is crucial to guide prognosis and clinical management of patients.

Diagnosis of a cancer subtype is reached with a tissue obtained by a biopsy that passes through a histological process, and it is subsequently evaluated by a pathologist [70]. This assessment includes a staging of the disease long before administering any treatment. For most cancers, the degree of the disease is determined, aiming to establish the abnormality level of the cancer cells. However, this complex process is dependent on the pathologist expertise, and yet the final level of quantification is too poor, despite many variables for the patient management depend on it, namely the recurrence risk, the response to the treatment, the prognosis and the disease evolution [23]. In low-grade (well-differentiated) cancers, cells look practically normal and they generally tend to grow slowly. In high-grade (poorly differentiated) cancers, cells are very different from normal ones. In this case, it is common to observe an accelerated growth of the tumor region and having a worse prognosis, which is why they may require different treatments [71]. Overall, cell types define treatment and prognosis, for example if the neoplasm exhibit mixtures of squamous cells and adenocarcinomas the diagnosis will be less accurate. In fact, the grade of poorly differentiated samples is more challenging and usually the sample shows SCC and ADC pattern combinations because the close- and mix- related histopathological features [27, 28].

Evaluation of the extent, aggressiveness and severity of cancer is still a very subjective task, yet it is at the very base of any decision. Pathologists usually quantify observations by subjectively assigning the percentage of a particular pattern, i.e., 25%, 50%, 75% or 100%. This analysis is complemented by descriptions of specific cellular and tissular features. This methodology of course hinders many subtle patterns and limits the possibility of performing quantitative population studies since these analyses are hardly comparable or reproducible. There exists therefore a basic need of developing interpretable scores for quantifying complicated histopathological patterns.

The present work presents a supervised subtype learning approach of histopathology tissue samples, using typical tissue samples, from ADC and SCC, selected by expert pathologists. The basic idea is to use a variational autoencoder to construct a latent space and to project onto it a selection of typical histopathological patterns which may describe objectively variations of these patterns. An autoencoder can be though of as a compression model which preserves relevant information of the original data when mapping the input to a space with a smaller dimension. A set of probability distribution functions approximate the structure of this space and generate not only observations but synthetic occurrences. Overall, the latent space representation of an autoencoder is linear since the projected sample may be propagated back to the original space of image samples and a linear combination of parameters produce a synthetic histology image sample.

Specifically, a set of typical patches of each pathology, patches representing ADC and SCC typical well-differentiated tissue patterns, were selected by an expert and projected to the latent space. The ADC and SCC images of tissue samples were observed to map at different

locations of the latent space and grouped around very different regions in a 2D projection. Finally, the differentiation grade for each NSCLC subtypes between the actual images of tissue samples of $256 \times 256 \times 3$ pixels and stratifies them according to typical features of each cancer subtype (ADC and SCC).

This paper is organized as follows: Section 2.2 introduces the methodology, Section 2.3 present the evaluation and results, finally, Section 2.4 present some conclusions.

2.2 Methodology

To the best of our knowledge, this is a first proposal to automatically establish a differentiation tumor tissue grade between two quite complex pathological entities (ADC and SCC). Variational Autoencoders (VAE) have shown to be useful for representation learning, data generation and dimensionality reduction [72, 73]. This type of autoencoder captures the invariant data relations and generates a linear representation of the high-dimensional data into a low-dimensional space, thereby facilitating formulation of simple metrics in the latent space [74, 75]. An aim of the present investigation was to construct a quantification of differentiation grade of tumoral tissue samples by estimating the proportion of each cancer subtype from an actual histopathology workflow. For doing so, a dataset of patients diagnosed with NSCLC (see Section 2.2.1) was used to train a VAE with a Gaussian prior (see Section 2.2.2). The latent space generated by the mixture of Gaussian distributions captures main invariants features of the subtypes of cancer by projecting each data into a space with smaller dimension. Relationships of this latent space are described by a logistic regression model (see Section 2.2.3) which estimates the staging and severity of the disease in a set of selected tissue samples from independent set of histopathology cases of both NSCLC subtypes.

2.2.1 NSCLC dataset

The dataset was then built from The Cancer Genome Atlas (TCGA) [76], a set of patients diagnosed with NSCLC, one slide per patient was selected from which several WSI regions of 1024×1024 pixels, diagnosed as typical of any of the two types of tumor (ADC and SCC), were randomly extracted. For each WSI region, squared patches (tissue samples) of 256×256 pixels from the tumor region were extracted at a magnification of $20\times$ with a microns per pixel ratio (MPP) of 0.5015 (see Figure **2-1**).

From the total of patches extracted automatically, typical patches of each class were manually selected with the help of pathologists. Poorly differentiated patches, with artifacts such as blurring or tissue with cellular absence, were discarded. In total, training set contains five cases from which 89 patches were extracted for the ADC subtype and five cases from which

78 patches were extracted for the SCC subtype. Testing set was built with ten different cases, five diagnosed with ADC and five with SCC from which 200 patches of ADC and 200 of SCC were extracted, as shown in the Table **2-1**.

Table 2-1: Distribution of cases and patches for ADC and SCC. Patches were selected by expert pathologists.

Subtype	Training		Testing	
	Cases	Patches	Cases	Patches
Adenocarcinoma	5	89	5	200
Squamous Cell Carcinoma	5	78	5	200
Total	10	167	10	400

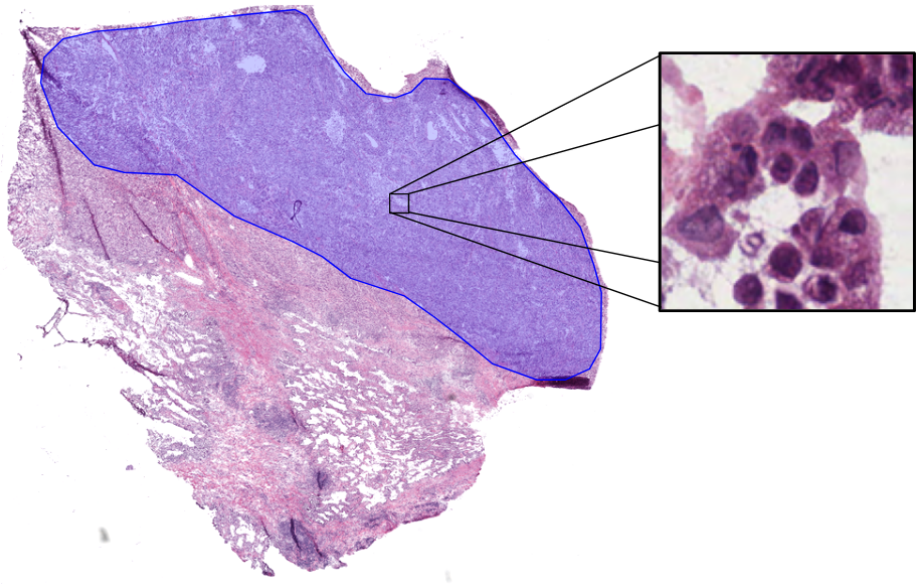

Figure 2-1: Sample image patch (tissue sample) of 256 × 256 pixels extracted from the annotated region (shaded region) of a histopathology case of (ADC). Own source.

2.2.2 Variational autoencoder

A VAE is basically a strategy which compresses information into a space with reduced dimension, the latent space, and recoveries the original data by reversing, as far as possible, the achieved compression. An interesting property of these methods is that they linearize data relations in such latent space [77, 78], making possible the development of enriched and more appropriate representations to attach and design quantitative and interpretable methods.

Basically, a VAE is a network composed of two coupled convolutional neural networks: an input which encodes the data to a lower dimensional space, the latent space, and an output which reconstructs original data from the probability distributions estimated in the latent space (see Figure **2-2**). Herein the encoder network consists of five convolutional blocks (convolution and downsampling), a deep fully-connected layer, two fully-connected layers in parallel and an output layer. Convolutional blocks reduce the input data dimension from $256 \times 256 \times 3$ to $8 \times 8 \times 1024$ pixels, while the deep fully connected layer extracts feature maps to be processed by two fully-connected layers in parallel, one in charge of estimating means and the other variances. These parameters are used by the output layer to construct 128 normal probability distributions, assuming the distribution of the latent space is approximated by a summation of normal distributions [79].

The Parzen estimator constructs an isotropic Gaussian for each projected data in the manifold and determines the density by observing the neighbors within the bounded region. The decoder network processes the latent space by sampling an instance from the mixture of Gaussian distributions. In this point, a fully connected layer interpolates the data to the dimensions of the encoder network ($8 \times 8 \times 1024$), followed by five convolutional blocks (convolution and upsampling) connected to this layer with the same number of parameters as the encoder network layers, i.e., the architecture of the encoder and decoder networks are the same and likewise the number of parameters. However, the parameters learned by both networks are not necessarily the same. The upsampling layers perform a data interpolation taking into account the nearest neighbor to scale up the data to the original dimension ($256 \times 256 \times 3$ pixels). The final result of the network is the reconstruction of the input data from the sampling performed in the latent space.

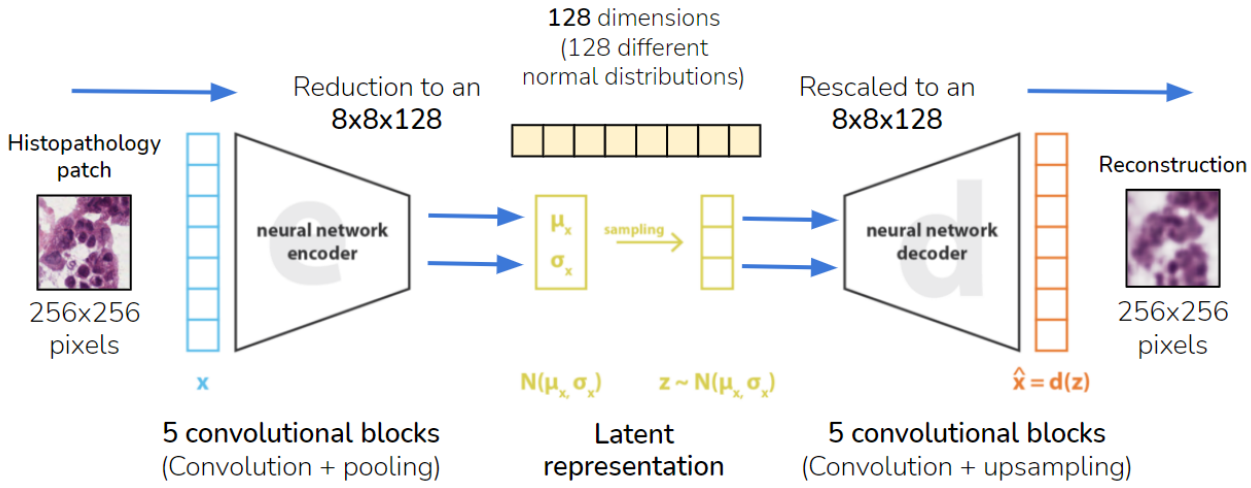

Figure 2-2: VAE architecture. The input patch is encoded to built a 128-dimensional latent space, where data is sampled and interpolated to reconstruct the patch to its original dimension. Adapted from Rocca (2019) [5].

Latent space

These VAE minimize the reconstruction error and capture main input data invariants in a latent space which is approximated by a set of probability distributions, making possible to generate not only observations but also synthetic examples. This latent space, after Parzen approximation, consists of a set of normal distributions, reason why the loss function is composed of two terms, a reconstruction term (in the decoder output layer) and a regularization term (in the encoder output layer) (see Figure **2-3**). The reconstruction term is the difference between the image reconstructed by the decoder network and the input image, while the regularization term is expressed as the Kulback-Leibler divergence between the data distribution in the latent space and a standard normal distribution. This term prevents the model from encoding data too far apart in the latent space and the generated probability distributions to overlap. By overlapping probability distributions, the generation of new data shares common features.

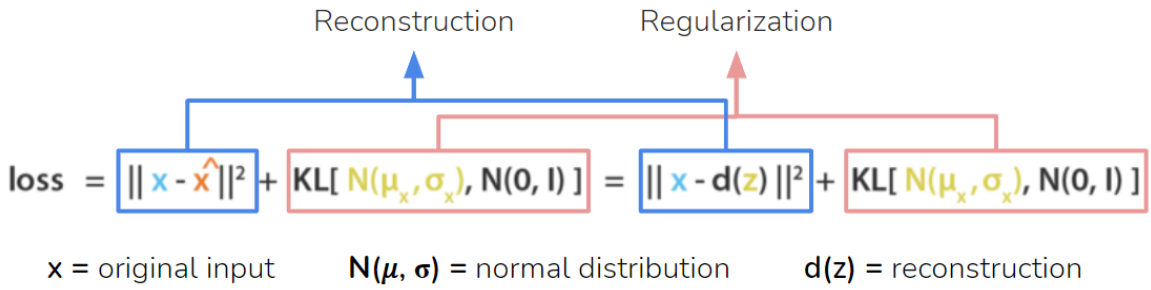

Figure 2-3: The loss function is composed of a reconstruction term (which minimizes the input-output difference between of the VAE) and a regularization term (makes the latent space more regular). Adapted from Rocca (2019) [5].

2.2.3 Latent space representation for cancer subtype differentiation

A logistic regression model with the Limited-memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) optimization method was implemented by taking as input the values generated by the probability distributions of the latent space using the linear representation of 128 values of means and variance learned by the VAE. LBFGS optimization stores only the latest updates of the second derivative of the matrix with gradient evaluations. The selected image patches of well-differentiated tissue samples of ADC and SCC are mapped to different regions in the latent space and a quantification of the differentiation grade of actual image patches of $256 \times 256 \times 3$ pixels are used to train the supervised method. Interestingly, the image samples with high-confidence estimation of the differentiation grade are regions with more typical patterns of a NSCLC subtype, either ADC or SCC, while patches with low-confidence estimation of the differentiation grade were projected to the regions in the latent space that exhibits complex mixtures of the two subtypes. This supervised subtype

differentiation learning model stratifies patches with different and complex visual patterns and histopathological characteristics of each NSCLC subtype and differentiation grade.

2.3 Results and discussion

2.3.1 Projections to the latent space

The two 128-dimensional latent spaces (mean and variance) generated by the VAE was projected into a two two-dimensional planes using the t-SNE (*t*-distributed Stochastic Neighbor Embedding) dimensionality reduction technique with Barnes-Hut approximation.

Figure **2-4** plots corresponding mean and variance values of the projected patches. As expected, the variance plots show better separability of the regions corresponding to the variability of histopathology visual features in the latent space per each NCSLC subtype, i.e. ADC (blue) or SCC (red). The areas with highest concentration of samples exhibit typical well-differentiation NCSLC subtype patterns (ADC or SCC), while values in the mixture regions, and fuzzy frontier, contain poorly differentiation of NCSLC subtypes for complex and more close proportions of the combinations of their histopathology tissue samples.

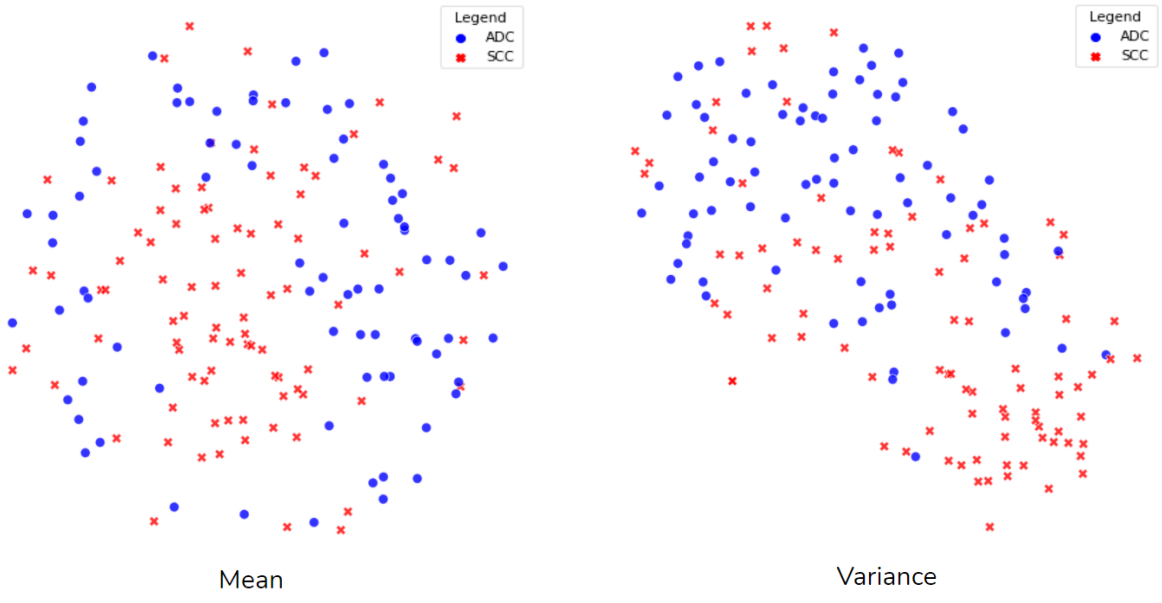

Figure 2-4: Graphs corresponding to the means and variances of the projected data in the latent space. Representative data for each NCSLC subtype are concentrated in different regions of the 2D-projected latent space. Own source.

2.3.2 Measuring differences between ADC and SCC

SCC is a neoplasia of epithelial origin, morphologically characterized by keratinization and/or intercellular bridges. ADC is also epithelium-originated, with varied architectural patterns, i.e., acinar, papillary, micropapillary and lepidic. However, tumors with a lower degree of differentiation show fuzzy or confused SCC and ADC visual features, case in which squamous or glandular identification is made by immuno-histology [80].

Qualitative results

Examples were selected for each NCSLC subtype and projected to the latent space (see Figure **2-5**). Patches A and B are examples of the ADC subtype, the first located in the region with highest ADC concentration corresponding to well-differentiated ADC region in the latent space, while the second at the poorly differentiated region with mixture of ADC and SCC. Patches C and D correspond to examples of the SCC subtype, the first located at the poorly differentiated region with mixture of ADC and SCC, it exhibits areas of keratinization, but also artifacts which simulate glandular patterns. Finally, patch D is located near the well-differentiated SCC region in the latent space, showing typical SCC patterns such as the intercellular bridges.

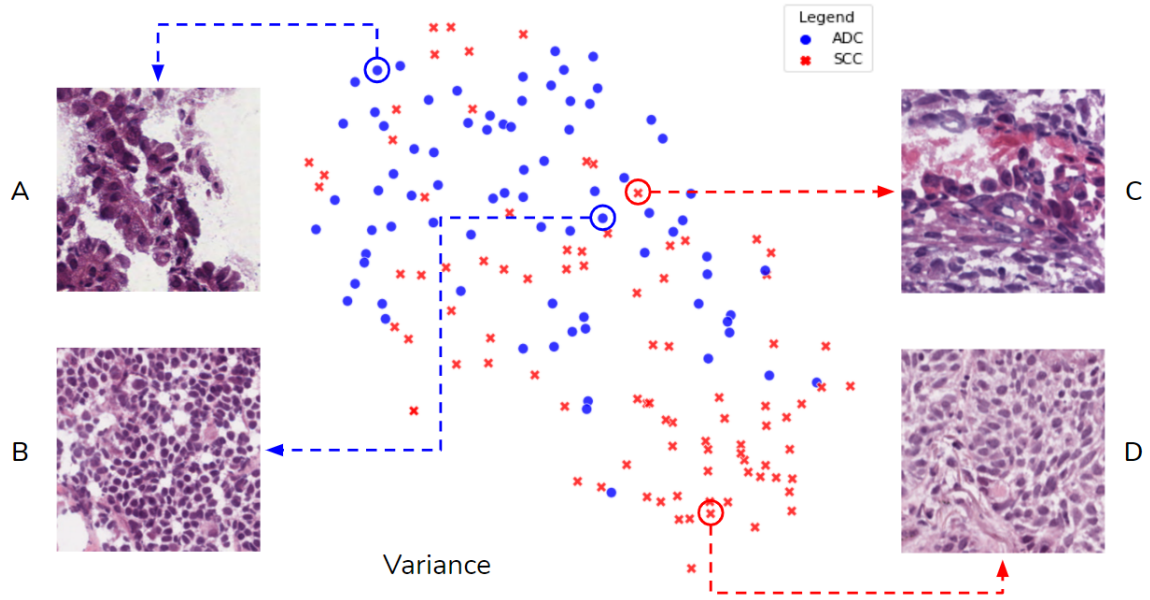

Figure 2-5: Projected ADC and SCC examples in a 2D-plane by t-SNE. Patches with typical characteristics are concentrated around the same latent space region projected while poorly differentiated patches are located in the mixture regions of the latent space projection. Own source.

Quantitative results

The validation results with the performance measure of the generalization capability of the presented approach is presented. A logistic regression model estimates the differentiation grade of NCSLC subtypes using the invariant patterns of the tissue samples. Several runs of random five cross-validation was performed selecting six cases (three per NCSLC subtype) with their corresponding tissue samples for training and four cases (two per NCSLC subtype) for validation to explore the regularization parameter of logistic regression model. Performance measures Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) were calculated. The best model in cross-validation was obtained using $C = 0.1$ reaching an average performance of $MAE = (0.072 \pm 0.0004)$ and $RMSE = (0.2654 \pm 0.0019)$. Thus, a logistic regression model was trained with the full training set of six cases with the best parameter in cross-validation step. Finally, using an independent testing data of lung cancer for evaluation, the model achieves a performance of $MAE = 0.2275$ and $RMSE = 0.477$.

Figure **2-6** shows some examples of the quantification of image patches to estimate the differentiation grade for each NCSLC subtype by the logistic regression model. The probability of belonging to one of the two classes is shown at the top of each example. Patches with SCC well-differentiated features are located at the extreme left, with a higher probability of belonging to the class.

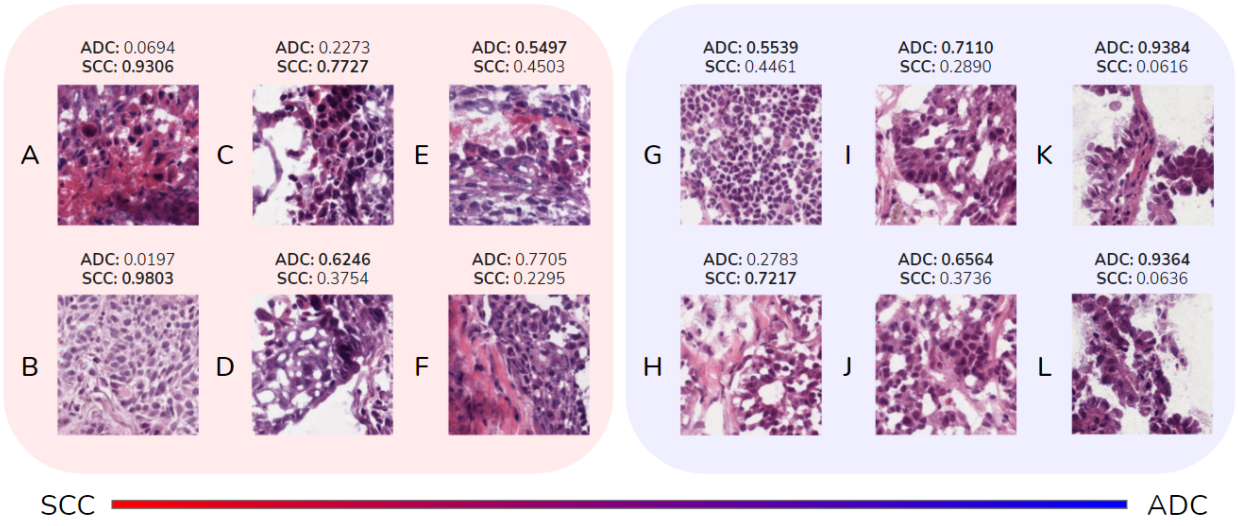

Figure 2-6: Examples of results obtained with the logistic regression model with the training set. The well-differentiated image patches of each NCSLC subtypes are located at the extremes, SCC on the left side (red) and ADC on the right side (blue), while the poorly differentiated patches are located in the center of the graph. Own source.

Pathologists analysis

Patches A and C show keratinization, while patch B exhibits intercellular bridges, typical of SCC with the highest value of SCC = 0.9803. Patches D and F, despite being SCC patches, show ADC features, namely pseudo-glandular spaces similar to the ADC acinar pattern, that explains the lower values of SCC. The cellular arrangement around the structures observed in patches I and J, correspond to the ADC class and the higher value of ADC. Patch F shows a poor differentiated pattern, the like-glandular structure together with the absence of SCC features with the corresponding lower value of SCC.

Features in patches E and G are unspecific and no classification was possible, as in patch G, where nuclei are small, crowded, with scant cytoplasm and with no structures of glandular lineage, represented in the closed valued for ADC = 0.5539 and SCC = 0.4461. On the other hand, patch E exhibits few areas of keratinization. Patches K and L, ADC patches, show well-differentiated features of ADC like rows of columnar cells, a central eosinophilic structure and nuclear polarization towards the periphery, in addition, patch K shows a lepidic pattern, which explains the higher values of ADC for both K and L patches. All these features suggest the adenocarcinoma subtype.

Finally, patch H, typical of the ADC class, contains a wide area of eosinophilic extracellular material that could simulate the presence of keratinization, a typical pattern of the SCC class, which corresponds to the wrong characterization by the model and it explains the lower value of 0.2783 for ADC.

2.4 Conclusions

This investigation explored a novel approach to quantify the differentiation subtype of non-small cell lung cancer lesions, using patterns learned by a variational autoencoder. The latent space generated by the VAE allows the development of enriched and more appropriate representations to attach and design quantitative and interpretable methods. A logistic regression model quantify the differentiation subtype of image patches with different histopathology visual patterns, characteristic of each NCSCLC subtype (ADC or SCC).

It was found that the representative well-differentiated patterns of each NCSCLC are concentrated in the same regions of the latent space with highest concentration of samples, and poorly differentiated image patches or with mixed histopathology features of both NCSCLC subtypes are concentrated in other mixture regions in the latent space. This early and promising results envision the possibility to quantify the proportion of complex tissue cancer subtypes thanks of a pathologist's data-addressed learning approach of a supervised subtype differentiation learning of non-small cell lung cancer to include reproducible and quantitative properties in computational pathology workflow.

3 Ensemble of image representations for ADC subtypes

This chapter presents a novel unsupervised computational approach based on an ensemble of tissue-specialized variational autoencoders, which were trained to differentiate the main histological ADC subtypes, and build an unsupervised embedded tissue-image representation. Each VAE encodes the information and specializes in building a representation of a sample of input images from the same tissue subtype. Subsequently, they are concatenated to build a final tissue descriptor based on the patterns learned by each VAE. This representation was used to train a Random Forest classifier of three lung adenocarcinoma histology subtypes (lepidic, papillary and solid), and a 2D-visually interpretable projection from the learned embedded representation. Experimental results achieve an average F-score of 0.72 ± 0.05 in the test dataset and a well-separated 2D visual mapping of tissue subtypes. This approach demonstrated that specialized models could be obtained with a small randomly selected data sample, and therefore, how the representation versatilely allows its use to distinguish among ADC histological subtypes with high throughput and low variance. A complete version of this chapter was submitted for presentation as a research article in the **20th IEEE International Symposium on Biomedical Imaging - ISBI 2023**, which will be held in Cartagena de Indias, Colombia, April 18-21, 2023.

3.1 Introduction

Lung adenocarcinoma (ADC) is one of the most frequent types of non-small cell lung cancer (NSCLC), whose incidence is estimated to be close to 40%; followed by squamous cell carcinoma (SCC), with an incidence around 30% of all cases [11, 12]. Several studies have demonstrated the importance of discriminating NSCLC subtypes [14] since available treatments are different depending on the histological subtype [15]. Particularly, ADC exhibits different patterns known as histological subtypes which include lepidic, acinar, papillary, micropapillary and solid [22], as shown in the Figure **3-1**. Each pattern has a known prognosis and aggressiveness [19]. Lepidic pattern is the least aggressive and has the best prognosis, acinar and papillary patterns are considered intermediate, and the micropapillary and solid patterns are the most aggressive and are generally associated with the worst prognosis.

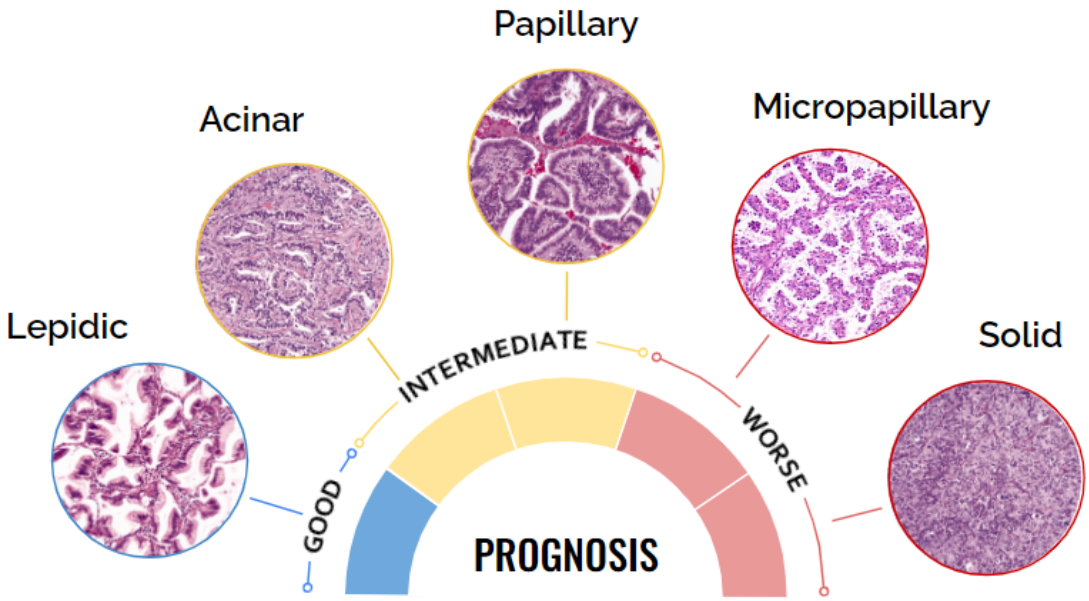

Figure 3-1: The five main histological subtypes of ADC correlate with its most common prognosis. Adapted from: Kuhn, E. (2018) [3].

Precise and early characterization of NSCLC is crucial to determine patient prognosis and survival [16]. However, the incidence of NSCLC subtypes varies widely, whereby there are some subtypes more common than others, and since about 80% of ADC cases contain a mixed spectrum of multiple histological patterns and unspecified tissues [27]. In addition, patient prognosis is estimated according to tumor grade, which is typically based on tissue architecture and cellular characteristics. Different from other types of cancer, e.g., breast or prostate cancer, in lung cancer (LCa) there is no standardized grading score system. Therefore, tumor differentiation and grading process is subjective by pathologists and currently determined on a spectrum going from well-differentiated (grade 1) to poorly-differentiated (grades 3 or 4) [18]. In fact, the qualitative criteria used to differentiate histological patterns tend to induce variability among pathologists [81, 20].

Previously, computational pathology approaches have reported the potential to learn typical patterns that characterize the NSCLC histological subtypes, as well as, to guide pathologists and oncologists in improving the accuracy of medical diagnosis [33]. For instance, Coudray *et al.* [14] developed a deep learning model for automatic analysis of NSCLC histological subtypes using data retrieved from The Cancer Genome Atlas (TCGA) and proprietary cohorts. This work demonstrated the capacity of a convolutional neural network (CNN) to support LCa diagnosis in difficult-to-diagnose cases. Likewise, Gertych *et al.* [19], implemented a CNN to characterize four ADC histological patterns, i.e., acinar, micropapillary, solid, and cribriform. The authors presented a comparison between results obtained with a CNN model and pathologist interpretation, showing how machine learning models like CNN

have the potential to distinguish ADC histological subtypes and the associated growth patterns. Similarly, Wei *et al.* also implemented a CNN to differentiate four ADC histological subtypes (lepidic, papillary, micropapillary and solid) and compared the results with pathologist annotations, concluding their model had the potential to classify the subtypes and assist pathologists in clinical practice by pre-scanning the image and highlighting regions of diagnostic interest. Also recent efforts have emerged, such as the challenge WSSS4LUAD [59], organized by The 5th International Symposium on Image Computing and Digital Medicine (ISICDM 2021), in which it was sought to characterize some types of tissue in ADC images. The designed computational models proved to be a feasible solution to replace the manual characterization of the tissues, being of great interest to reduce annotation efforts. In addition to the multiple applications, computational models in digital pathology are including pathologists and oncologists as guarantors of their operation and as support to develop more specialized models that allows to understand the nature of pathologies [20, 59].

This work introduces a computational pathology approach based on an ensemble of tissue-specialized variational autoencoders, each of them trained with the associated histopathology ADC subtype, i.e., lepidic, papillary and solid. This approach aims to build an unsupervised embedded tissue-image representation used to: 1) train a Random Forest tissue classifier of ADC subtypes, and 2) construct a 2D projection that is visually interpretable of tumor tissue sample distribution of ADC subtypes in the embedded space. Our proposed approach of unsupervised embedded tissue-image representation allows a good histopathology ADC subtypes differentiation, and a semantic spatial distribution of tissues into embedded space, placing typical isolated histopathology patterns of ADC subtypes (e.g. lepidic and solid) in the periphery, and untypical and mixtures of histopathology patterns (e.g. papillary) in the central zone. These two characteristics could help pathologist by providing more quantitative, objective and interpretable tools of computational pathology for diagnosis support in the more challenging and relevant tumor differentiation among histopathology ADC subtypes.

3.2 Methodology

3.2.1 ADC lung cancer dataset

A total of 41 cases were retrieved from The Cancer Genome Atlas (TCGA-LUAD) database. These cases correspond to Formalin-Fixed, Paraffin-Embedded (FFPE) biopsies of three of the main histological subtypes of ADC, distributed as: 16 lepidic, 20 papillary and 5 solid; representing one histological subtype per level of aggressiveness.

3.2.2 Preprocessing

As illustrated in Figure **3-2**, square patches of 256 × 256 pixels were extracted at 10× of magnification from tumor regions of each whole-slide image (WSI). According to Coudray [14], those values of patch size and magnification are the appropriate ones to identify characteristic patterns that allow differentiation of histological subtypes of ADC on WSI. For each square patch, a color deconvolution was performed to separate the hematoxylin and eosin stains, and only the hematoxylin channel was selected to be used because of better nuclei visualization. Finally, hematoxylin-channel patches were downscaled to a size of 128 × 128 pixels, in order to reduce the computational cost necessary for the training and testing of machine learning models.

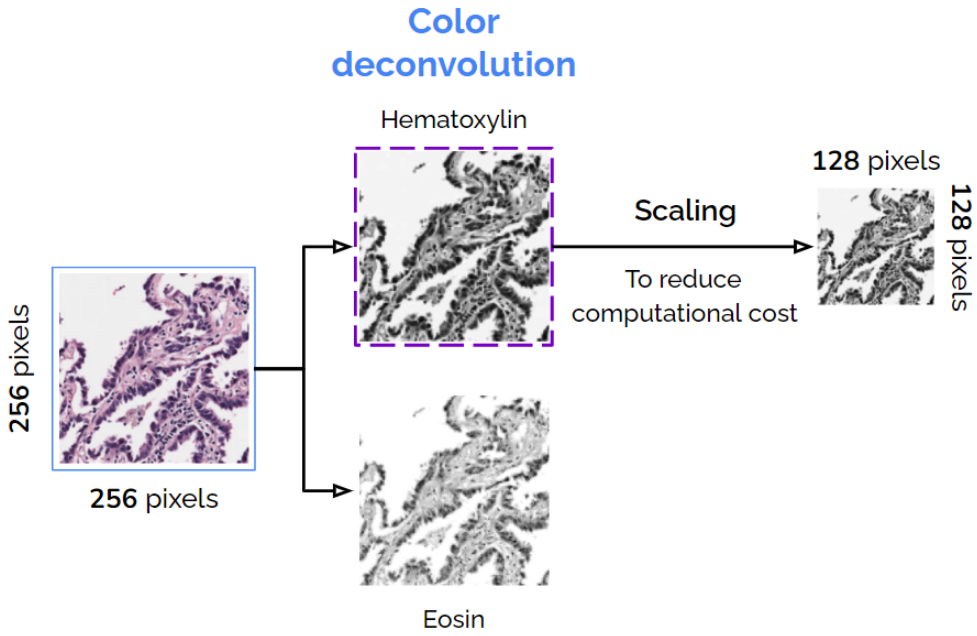

Figure 3-2: Preprocessing of patches extracted from tumor regions. Color deconvolution was performed to obtain hematoxylin and eosin channels, and then hematoxylin patches were downscaled to reduce computational cost in later steps. Own source.

3.2.3 Variational Autoencoder

An ensemble of tissue-specialized encoder-decoder architecture was designed, specifically based on Variational Autoencoders (VAE), i.e., a learned representation of each histological subtype of ADC is obtained per VAE. A VAE is a unsupervised learning model composed of two sections, encoder and decoder, for an embedded representation of mixture of normal distributions (see Figure **3-3**). On the one hand, the encoder reduces input dimensionality and generates a linear representation by learning a set of probability distributions, usually a mixture of normal distributions with mean and variance values. The mixture of normal

distributions are an approximation to the real data distribution, and their sampling allows reconstructing the input data, as well as generating new synthetic data. On the other hand, the decoder performs an upsampling of the samples extracted from the mixture of distributions to the original size of the input data. In summary, the purpose of the encoder is to generate a compressed and embedded representation of the input data as a mixture of probability distributions, while the decoder reconstructs the input image.

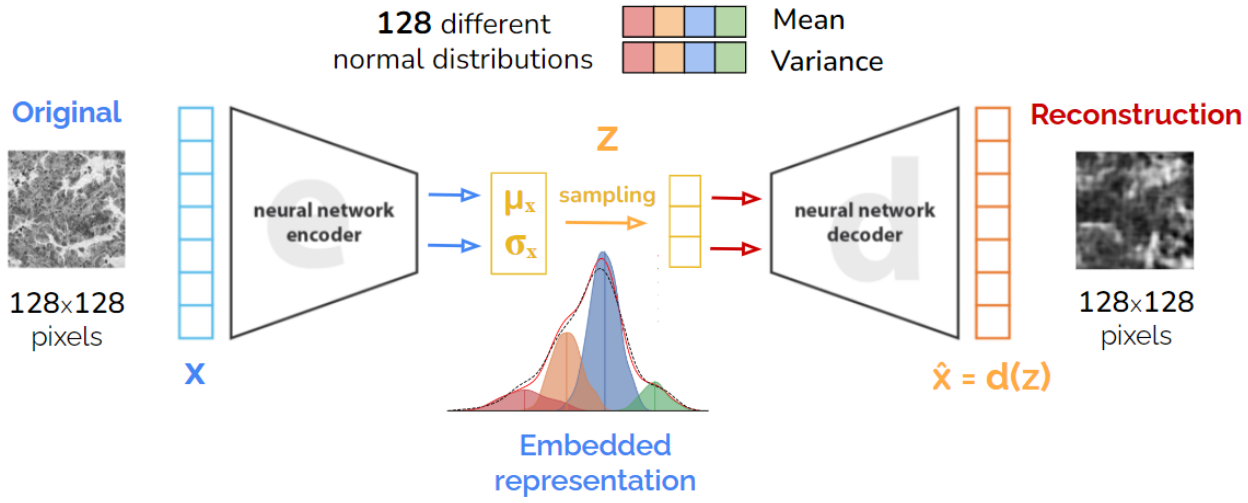

Figure 3-3: VAE architecture. Encoder reduces the input dimensionality from 128 × 128 pixels to a vector of 128 values, i.e., 128 normal probability distributions. Then, decoder upsamples these values to the original size, which corresponds to the reconstruction. Adapted from Rocca (2019) [5].

In this way, the loss function of a VAE architecture is composed of two terms, reconstruction and regularization:

$$\text{Loss} = \underbrace{||x - \hat{x}||^2}_{\text{Reconstruction term}} + \underbrace{\text{KL}[\text{N}(\mu_x, \sigma_x), \text{N}(0,1)]}_{\text{Regularization term}} \tag{3-1}$$

where, the reconstruction term is the difference between the model output $\hat{x}$ and its input x, and the regularization term is the Kullback-Leibler divergence KL between the estimated probability distribution of the data $\text{N}(\mu_x, \sigma_x)$ and a reference distribution $\text{N}(0,1)$ (standard normal).

3.2.4 Experimental setup

The ADC lung cancer dataset was divided into two parts (see Table **3-1**): 80% of cases of each histological subtype, that is, 13 lepidic, 16 papillary and 4 solid compose the training set (33 cases in total); and the remaining 20%, that is, 3 lepidic, 4 papillary and 1 solid

constitute the test set (8 cases in total). A small set of patches was randomly selected for each training case. Lepidic dataset is composed of 494 patches (38 per case), papillary with 496 patches (31 per case), and solid with 500 patches (125 per case), and no additional data augmentation process was performed. It is worth mentioning that the number of patches per case is not balanced between classes, since the main intention of this work is that each model specializes in one of the three selected histological ADC subtypes, so only a case balance is necessary. Additionally, the test dataset contains 600 lepidic patches (200 per case), 672 papillary patches (168 per case), and 720 solid patches.

Table 3-1: Distribution of cases and selected patches of the three main histological subtypes of ADC, one subtype for each level of aggressiveness.

Subtype	**Training**		**Testing**	
	Cases	**Patches**	**Cases**	**Patches**
Lepidic	13	494	3	600
Papillary	16	496	4	672
Solid	4	500	1	720
Total	33	1.490	8	1.992

A VAE was trained for each histological ADC subtype in order to specialize each network in differentiating tissue patterns of the subtype with which it was trained. Encoder and decoder are composed of 5 blocks, each block has a convolution layer, a L1 normalization and a max-pooling layer. Architecture and hyperparameters of each of the three models are the same. Models were trained in 500 epochs with 128 dimensions at its bottleneck, 128 probability distributions, and the reconstruction term in the loss function was regularized to obtain a better representation of the data, as described in *Higgins 2017* [66].

3.2.5 Local tissue representation

Training patches of each histological subtype were projected onto their corresponding VAE to obtain a feature vector of 128 values per patch, corresponding to sampled values of the 128 bottleneck probability distributions. Each VAE is specialized in compressing the information for each histological subtype. The mean and variance vectors allow the construction of a final vector, a mixture of probability distributions, where each position is the sampled value of a normal probability distribution, as shown in Figure **3-4**.

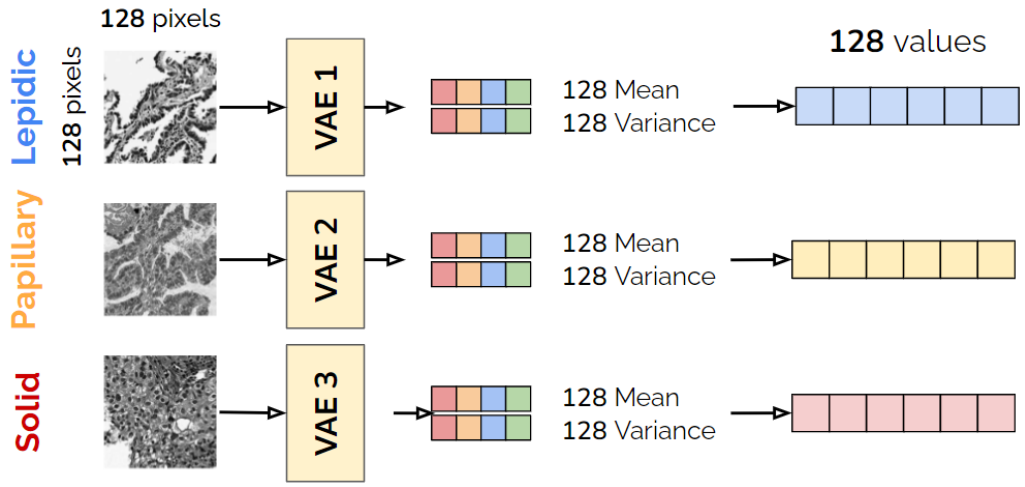

Figure 3-4: Embeddings for each histological subtype of ADC. Own source.

Subsequently, patches of each histological subtype were cross-projected onto the VAE corresponding to the remaining subtypes to obtain the whole tissue-based feature vectors of each patch based on the representation spaces of each histological ADC subtype. Finally, individual representations of 128 values were concatenated in a final vector of 384 features, as shown in the Figure **3-5**. In summary, each tissue patch is expressed as the concatenated representation of all selected histological ADC subtypes.

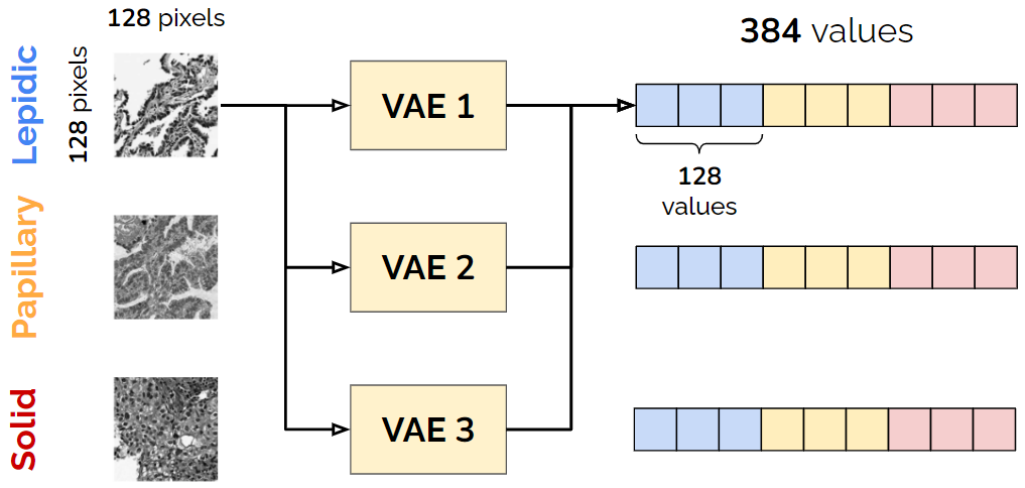

Figure 3-5: Concatenated representation of features. Each vector is composed of the individual representations of each tissue-specialized VAE. Own source.

3.3 Results and discussion

3.3.1 Qualitative results

In order to have a visual and interpretable representation of the data, a dimensionality reduction process was performed using a Uniform Manifold Approximation and Projection (UMAP) [82] with Euclidean metric, thus reducing the 384 concatenated dimensions per patch to two dimensions (2D), and subsequently plotted, as shown in Figure **3-6**.

It can be seen that patches with typical patterns of each histological subtype are located on

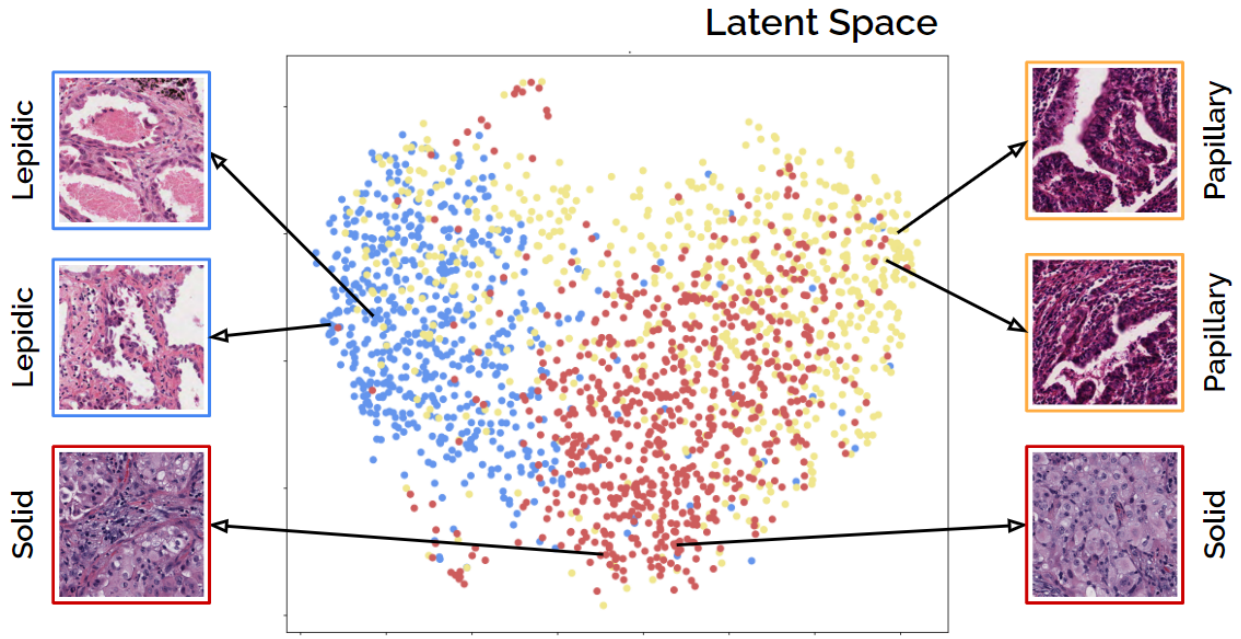

Figure 3-6: Representation space. Each point on the space represents an input patch. Own source.

the periphery of the space, forming visually separable regions. In turn, patches that contain mixtures of patterns are located in the central zone, which shows that the constructed representation is able to separate the relevant concepts using only the information obtained at the patch level, without any kind of supervision.

3.3.2 Quantitative results

Two different classification models were trained as a way to validate the discriminating capability of the representations learned by the VAEs: Support Vector Machines (SVM) with linear kernel and radial kernel basis function (RBF), and Random Forest. In each case, an random grid search exploration of parameters was carried out until the best model was obtained. The results are shown in Table **3-2**, and additionally the confusion matrices for

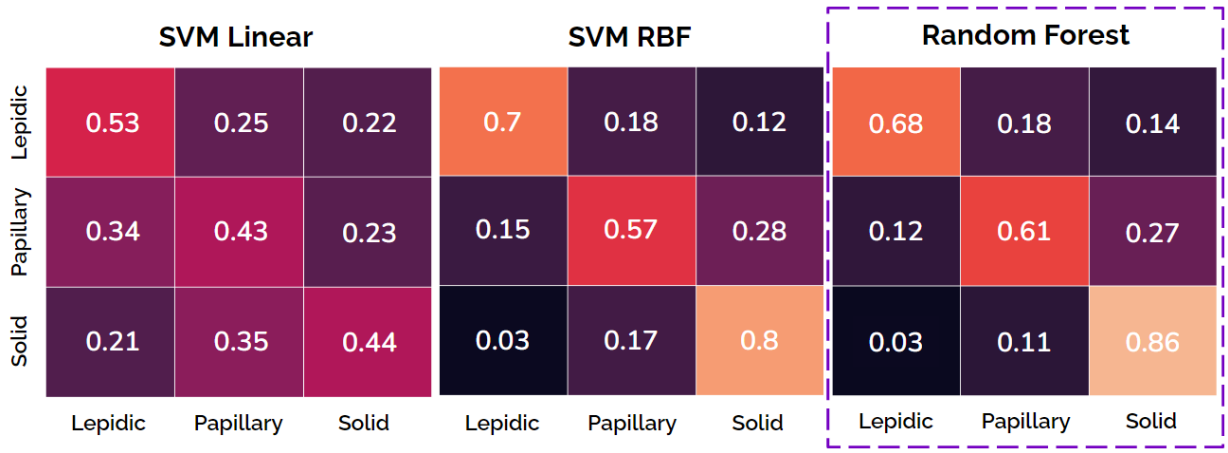

Figure 3-7: Confusion matrices for each of the classification models. Own source.

the three cases are presented (Figure **3-7**), where it is evident that the model that obtained the best performance was Random Forest, with an F1-Score of 0.72 ± 0.05.

Table 3-2: Quantitative performance results. A Random Forest classifier with 500 decision trees and a depth of 7 obtained the best performance.

Model	Accuracy	Precision	Recall	F1-Score
SVM Linear	0.46 ± 0.02	0.47 ± 0.04	0.47 ± 0.05	0.46 ± 0.03
SVM RBF	0.69 ± 0.04	0.69 ± 0.06	0.69 ± 0.09	0.69 ± 0.07
Random Forest	$\mathbf{0.72 \pm 0.03}$	$\mathbf{0.73 \pm 0.05}$	$\mathbf{0.72 \pm 0.11}$	$\mathbf{0.72 \pm 0.05}$

3.4 Conclusions

A novel approach, based on an ensemble of VAE learned representations was proposed to differentiate the main histological ADC subtypes. Each VAE encodes the information and specializes in building a representation of a sample of input images from the same tissue subtype. Subsequently, they are concatenated to build a final tissue descriptor based on the patterns learned by each VAE. This approach demonstrated that specialized models could be obtained with a small randomly selected data sample, and therefore, how the representation versatilely allows its use to distinguish among ADC histological subtypes with high throughput and low variance. In addition, the presented approach is discriminating enough to show each tissue subtype in a two-dimensional projection for easy visual interpretation, identifying the different patterns that characterize tissues according to their interpretation complexity, either by typicity or mixture of histological patterns. Future work includes to extend our approach with more histopathology ADC subtypes and evaluate its potential to design quantitative image-based biomarkers of tumor grading and prognosis.

4 Conclusions and perspectives

4.1 Conclusions

This investigation explored a novel approach to quantify the differentiation grade of non-small cell lung cancer subtypes, using patterns learned by a variational autocoder. The latent space generated by the VAE allows the development of enriched representations that are more suitable for linking and designing quantitative and interpretable methods.

Well-differentiated patterns representative of each NSCLC subtype were found to be concentrated in the same regions of the latent space with the highest sample concentration, and poorly differentiated or mixed histopathological imaging patches of both NSCLC subtypes were concentrated in other mixed regions in the latent space.

In addition, the representations learned by each VAE can be specialized for each histological subtype from a small sample of selected data, and by performing an ensemble of these projections on the spaces learned by the VAEs, also it is possible to distinguish in a versatile way among the histological subtypes of ADC with high throughput and low variance. Furthermore, this approach is discriminating enough to display each tissue subtype in a two-dimensional projection for easy visual interpretation, identifying the different patterns that characterize the tissues based on their complexity of interpretation, either by typicality or a mixture of histological patterns.

4.2 Perspectives

In this work, a data-driven model was designed to characterize the patterns that define the histological subtypes of lung cancer. In the construction of the work, different approaches were tested using few data motivated by the difficulty of acquiring large amounts of characterized data in the biomedical context. However, these challenges are still present, so they are the starting point for new directions for future research work:

- Extend the proposed approach with the full set of histological subtypes that characterize ADC and assess its potential to design quantitative image-based biomarkers for tumor classification and prognosis.

- A comprehensive solution for the pathology workflow. Since the starting point in the diagnosis of lung cancer are radiology images, more commonly computed tomography images, to propose a solution that tries to combine the information from these types of images with the information learned from histopathology images. It could open the way to constructing enriched metrics that make it possible to characterize the pathology in a more efficient manner.

Bibliography

[1] V. Moctezuma and Z. Patiño, "Cáncer de pulmón," *Anales de Radiología México*, vol. 8, pp. 33–45, 2009.

[2] E. Bender, "Epidemiology: The dominant malignancy," *Nature*, vol. 513, pp. 52–53, 2014.

[3] E. Kuhn, P. Morbini, and et. al., "Adenocarcinoma classification: patterns and prognosis," *Pathologica*, vol. 110, pp. 5–11, 2018.

[4] C. Underwood, A. Musick, and C. Glass, "Adenocarcinoma overview." https://www.pathologyoutlines.com/topic/lungtumoradenocarcinoma.html, 2019. Online; accessed Jun 2022.

[5] J. Rocca, "Understanding variational autoencoders (vaes)." https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73, 2019. Online; accessed Nov 2022.

[6] A. Moreira, P. Ocampo, Y. Xia, and et. al., "A grading system for invasive pulmonary adenocarcinoma: A proposal from the international association for the study of lung cancer pathology committee," *J Thorac Oncol*, vol. 15(10), pp. 1599–1610, 2020.

[7] M. Amin and et. al., *AJCC Cancer Staging Manual. 8th ed.* New York, NY.: Springer, 2017.

[8] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries.," *CA: a cancer journal for clinicians*, vol. 71(3), p. 209–249, 2021.

[9] M. Šutić, A. Vukić, J. Baranašić, A. Försti, F. Džubur, M. Samaržija, and et al., "Diagnostic, predictive, and prognostic biomarkers in non-small cell lung cancer (nsclc) management," *J Pers Med.*, vol. 11, p. 1102, 2021.

[10] R. Wilson, A. Ryerson, K. Zhang, and D. X., "Relative survival analysis using the centers for disease control and prevention's national program of cancer registries surveillance system data, 2000-2007," *Journal of registry management*, vol. 41, p. 72, 2014.

[11] S. Blandin Knight, P. Crosbie, and et. al., "Progress and prospects of early detection in lung cancer," *Open biology*, vol. 7, 2017.

[12] Z. Wang, M. Li, and et. al., "Clinical and radiological characteristics of central pulmonary adenocarcinoma: a comparison with central squamous cell carcinoma and small cell lung cancer and the impact on treatment response," *OncoTargets and therapy 2018*, vol. 11, 2018.

[13] L. Osmani, F. Askin, E. Gabrielson, and Q. Li, "Current who guidelines and the critical role of immunohistochemical markers in the subclassification of non-small cell lung carcinoma (nsclc): Moving from targeted therapy to immunotherapy," *Semin Cancer Biol*, vol. 52, p. 103–9, 2018.

[14] N. Coudray and et. al., "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," *Nature Medicine*, vol. 24, pp. 1559–1567, 2018.

[15] V. S. Viswanathan, P. Toro, and et. al., "The state of the art for artificial intelligence in lung digital pathology," *The Journal of pathology*, vol. Advance online publication, 2022.

[16] C. Zappa and S. A. Mousa, "Non-small cell lung cancer: current treatment and future advances," *Translational lung cancer research*, vol. 5, pp. 288–300, 2016.

[17] P. Bolaños and C. Chacón, "Escala patológica de gleason para el cáncer de prostata y sus modificaciones," *Medicina Legal de Costa Rica*, vol. 34, pp. 237–243, 2017.

[18] W. Travis, E. Brambilla, and K. Geisinger, "Histological grading in lung cancer: one system for all or separate systems for each histological type?," *European Respiratory Journal*, vol. 47, pp. 720–723, 2016.

[19] A. Gertych, Z. Swiderska-Chadaj, and et. al., "Convolutional neural networks can accurately distinguish four histologic growth patterns of lung adenocarcinoma in digital slides," *Sci Rep*, vol. 9, p. 1483, 2019.

[20] J. Wei, L. Tafe, and Y. e. a. Linnik, "Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks," *Sci Rep*, vol. 9, p. 3358, 2019.

[21] M. Davidson, A. Gazdar, and B. Clarke, "The pivotal role of pathology in the management of lung cancer," *J Thorac Dis*, vol. 5, 2013.

[22] P. Russell, Z. Wainer, G. Wright, M. Daniels, M. Conron, and R. Williams, "Does lung adenocarcinoma subtype predict patient survival?: A clinicopathologic study based on the new international association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary lung adenocarcinoma classification," *J Thorac Oncol*, vol. 6, pp. 1496–1504, 2011.

[23] S. Wang, A. Chen, L. Yang, and et. al., "Comprehensive analysis of lung cancer pathology images to discover tumor shape and boundary features that predict survival outcome," *Sci Rep*, vol. 8, p. 10393, 2018.

[24] M. Van Bockstal, M. Berlière, F. Duhoux, and C. Galant, "Interobserver variability in ductal carcinoma in situ of the breast," *J Clin Pathol*, vol. 154, pp. 596–609, 2020.

[25] S. Zachara-Szczakowski, T. Verdun, and A. Churg, "Accuracy of classifying poorly differentiated non–small cell lung carcinoma biopsies with commonly used lung carcinoma markers," *Hum Pathol*, vol. 46, pp. 776–782, 2015.

[26] M. Connor, G. Canal, and C. Rozell, "Classification and pathology of lung cancer," *Surgical Oncology Clinics*, vol. 25, pp. 447–468, 2016.

[27] E. Ruffini, O. Rena, and et. al., "Lung tumors with mixed histologic pattern. clinicopathologic characteristics and prognostic significance," *European Journal of Cardio-Thoracic Surgery*, vol. 22, p. 701–707, 2002.

[28] J. Qin and H. Lu, "Combined small-cell lung carcinoma," *OncoTargets and therapy*, vol. 11, p. 3505–3511, 2018.

[29] W. D. Travis, E. Brambilla, M. Noguchi, A. G. Nicholson, K. R. Geisinger, Y. Yatabe, D. G. Beer, C. A. Powell, G. J. Riely, P. E. Van Schil, K. Garg, J. H. Austin, H. Asamura, V. W. Rusch, F. R. Hirsch, G. Scagliotti, T. Mitsudomi, R. M. Huber, Y. Ishikawa, J. Jett, and D. Yankelewitz, "International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma," *Journal of thoracic oncology: official publication of the International Association for the Study of Lung Cancer*, vol. 6, pp. 244–285, 2011.

[30] L. Gengpeng, L. Hui, K. Jianyi, T. Kejing, G. Yubiao, H. Anjia, and X. Canmao, "Acinar-predominant pattern correlates with poorer prognosis in invasive mucinous adenocarcinoma of the lung," *American Journal of Clinical Pathology*, vol. 149, p. 373–378, 2018.

[31] K. Kadota, Y. C. Yeh, C. S. Sima, V. W. Rusch, A. L. Moreira, P. S. Adusumilli, and W. D. Travis, "The cribriform pattern identifies a subset of acinar predominant tumors with poor prognosis in patients with stage i lung adenocarcinoma: a conceptual

proposal to classify cribriform predominant tumors as a distinct histologic subtype," *Modern pathology : an official journal of the United States and Canadian Academy of Pathology*, vol. 27, p. 690–700, 2014.

[32] T. Kinno, K. Tsuta, K. Shiraishi, T. Mizukami, M. Suzuki, A. Yoshida, K. Suzuki, H. Asamura, K. Furuta, T. Kohno, and R. Kushima, "Clinicopathological features of nonsmall cell lung carcinomas with braf mutations," *Ann Oncol*, vol. 25, p. 138–42, 2014.

[33] K. Bera, K. Schalper, and et. al., "Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology," *Nat Rev Clin Oncol*, vol. 16, pp. 703–715, 2019.

[34] M. Gurcan, L. Boucheron, A. Can, A. Madabhushi, N. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE Rev Biomed Eng*, vol. 2, pp. 147–171, 2009.

[35] L. Pantanowitz, D. Hartman, Y. Qi, E. Cho, B. Suh, K. Paeng, R. Dhir, P. Michelow, S. Hazelhurst, S. Song, and S. Cho, "Accuracy and efficiency of an artificial intelligence tool when counting breast mitoses," *Diagn Pathol*, vol. 15, p. 80, 2020.

[36] R. Ding, P. Prasanna, G. Corredor, and et al., "Image analysis reveals molecularly distinct patterns of tils in nsclc associated with treatment outcome," *npj Precis. Onc*, vol. 6, p. 33, 2022.

[37] A. Ibrahim, P. Gamble, R. Jaroensri, M. Abdelsamea, C. Mermel, P.-H. Chen, and E. Rakha, "Artificial intelligence in digital breast pathology: Techniques and applications," *The Breast*, vol. 49, pp. 267–273, 2020.

[38] M. Yousif, P. van Diest, A. Laurinavicius, D. Rimm, J. van der Laak, A. Madabhushi, S. Schnitt, and L. Pantanowitz, "Artificial intelligence applied to breast pathology," *Virchows Arch*, 2021.

[39] D. Van Booven, M. Kuchakulla, R. Pai, F. Frech, R. Ramasahayam, P. Reddy, M. Parmar, R. Ramasamy, and H. Arora, "A systematic review of artificial intelligence in prostate cancer," *Res Rep Urol*, vol. 13, pp. 31–39, 2021.

[40] O. Tătaru, M. Vartolomei, J. Rassweiler, O. Virgil, G. Lucarelli, F. Porpiglia, D. Amparore, M. Manfredi, G. Carrieri, U. Falagario, D. Terracciano, O. de Cobelli, G. Busetto, F. Giudice, and M. Ferro, "Artificial intelligence and machine learning in prostate cancer patient management—current trends and future perspectives," *Res Rep Urol*, vol. 11, p. 354, 2021.

[41] G. Litjens, C. Sanchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, and et al, "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," *Scientific Reports*, vol. 6, 2016.

[42] B. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling, "Rotation equivariant cnns for digital pathology," *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, vol. 11071, 2018.

[43] S. Mukhopadhyay, M. Feldman, E. Abels, R. Ashfaq, S. Beltaifa, N. Cacciabeve, H. Cathro, L. Cheng, K. Cooper, G. Dickey, R. Gill, R. Heaton, R. Kerstens, G. Lindberg, R. Malhotra, J. Mandell, E. Manlucu, A. Mills, S. Mills, C. Moskaluk, M. Nelis, D. Patil, C. Przybycin, J. Reynolds, B. Rubin, M. Saboorian, M. Salicru, M. Samols, C. Sturgis, K. Turner, M. Wick, J. Yoon, P. Zhao, and C. Taylor, "Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: A multicenter blinded randomized noninferiority study of 1992 cases (pivotal study)," *Am J Surg Pathol*, vol. 42, pp. 39–52, 2018.

[44] Z. Ahmad, S. Rahim, M. Zubair, and J. Abdul-Ghafar, "Artificial intelligence (ai) in medicine, current applications and future role with special emphasis on its potential and promise in pathology: present and future impact, obstacles including costs and acceptance among pathologists, practical and philosophical considerations. a comprehensive review," *Diagn Pathol*, vol. 16, p. 24, 2021.

[45] A. Le Page, E. Ballot, C. Truntzer, V. Derangère, A. Ilie, D. Rageot, F. Bibeau, and F. Ghiringhelli, "Using a convolutional neural network for classification of squamous and non-squamous non-small cell lung cancer based on diagnostic histopathology hes images," *Sci Rep*, vol. 11, p. 23912, 2021.

[46] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015.

[47] N. Punn and S. Agarwal, "Modality specific u-net variants for biomedical image segmentation: a survey," *Artif Intell Rev*, 2022.

[48] H. Tizhoosh and L. Pantanowitz, "Artificial intelligence and digital pathology: Challenges and opportunities," *J Pathol Inform*, vol. 9, p. 38, 2022.

[49] R. Baeza-Yates and Z. Liaghat, "Quality-efficiency trade-offs in machine learning for text processing," *IEEE International Conference on Big Data (Big Data)*, pp. 897–904, 2017.

[50] K. Huang, L. Yang, Y. Wang, L. Huang, X. Zhou, and W. Zhang, "Identification of non-small-cell lung cancer subtypes by unsupervised clustering of ct image features

with distinct prognoses and gene pathway activities," *Biomedical Signal Processing and Control*, vol. 76, 2022.

[51] C. Weis, K. Weihrauch, K. Kriegsmann, and M. Kriegsmann, "Unsupervised segmentation in nsclc: How to map the output of unsupervised segmentation to meaningful histological labels by linear combination?," *Appl. Sci.*, vol. 12, p. 3718, 2022.

[52] C. Lynch, V. van Berkel, and H. Frieboes, "Application of unsupervised analysis techniques to lung cancer patient data," *PloS one*, vol. 12, p. 9, 2017.

[53] Y. Gao, Y. Ding, W. Xiao, Z. Yao, X. Zhou, X. Sui, Y. Zhao, and Y. Zheng, "A semi-supervised learning framework for micropapillary adenocarcinoma detection," *International journal of computer assisted radiology and surgery*, vol. 17, pp. 639–648, 2022.

[54] K. Shak, M. Al-Shabi, A. Liew, B. Lan, W. Leong, N. Kh, and M. Tan, "A new semi-supervised self-training method for lung cancer prediction," *arXiv:2012.09472*, 2020.

[55] R. Chen, C. Chen, Y. Li, T. Chen, A. Trister, R. Krishnan, and F. Mahmood, "Scaling vision transformers to gigapixel images via hierarchical self-supervised learning," *arXiv:2206.02647*, 2022.

[56] V. Thanh-Hung, L. Guee-Samg, Y. Hyung-Jeong, K. Sae-Ryung, O. In-Jae, , and K. Soo-Hyung, "Multi-task with variational autoencoder for lung cancer prognosis on clinical data," *In The 9th International Conference on Smart Media and Applications (SMA 2020)*, pp. 234–237, 2020.

[57] S. Wang, M. Zhou, Z. Liu, Z. Liu, D. Gu, Y. Zang, D. Dong, O. Gevaert, and J. Tian, "Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation," *Medical image analysis*, vol. 40, pp. 172–183, 2017.

[58] K.-H. Yu, F. Wang, G. Berry, C. Ré, R. Altman, M. Snyder, and I. Kohane, "Classifying non-small cell lung cancer types and transcriptomic subtypes using convolutional neural networks," *J Am Med Inform Assoc*, vol. 27, pp. 757–769, 2020.

[59] C. Han and et. al., "Wsss4luad: Grand challenge on weakly-supervised tissue semantic segmentation for lung adenocarcinoma," in *arXiv*, 2022.

[60] M. for Primary Industries, "University of louisville health adopts paige ai-enabled cancer detection software for enhanced cancer detection." https://www.businesswire.com/news/home/20211215005218/en/University-of-Louisville-Health-Adopts-Paige-AI-enabled-Cancer-Detection-Software-for-Enhanced-Cancer-Detection, 2021. Online; accessed May 2022.

[61] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. Madai, "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," *BMC Med Inform Decis Mak*, vol. 20, p. 310, 2020.

[62] A. Madabhushi and G. Lee, "Image analysis and machine learning in digital pathology: Challenges and opportunities," *Medical image analysis*, vol. 33, pp. 170–175, 2016.

[63] U. Michelucci, "An introduction to autoencoders," 2022.

[64] A. Fuxjaeger and V. Belle, "Logical interpretations of autoencoders," 2019.

[65] D. Kingma and M. Welling, "An introduction to variational autoencoders," *Foundations and Trends in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.

[66] I. Higgins, L. Matthey, and et. al., "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *International Conference on Learning Representations*, 2017.

[67] D. Kingma and M. Welling, "Auto-encoding variational bayes," 2013.

[68] F. Cano, C. Alvarez-Jimenez, D. Becerra, A. Siabatto, A. Cruz-Roa, and E. Romero, "A supervised subtype differentiation learning for building invariant features of non-small cell lung cancer in a latent space of a variational autoencoder," in *Proc. SPIE 12088, 17th International Symposium on Medical Information Processing and Analysis*, 2021.

[69] F. Bray, J. Ferlay, I. Soerjomataram, and et. al., "Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians 2018*, vol. 68, pp. 394–424, 2018.

[70] W. Wu, C. Parmar, P. Grossmann, J. Quackenbush, P. Lambin, J. Bussink, R. Mak, and H. Aerts, "Exploratory study to identify radiomics classifiers for lung cancer histology," *Front Oncol.*, vol. 6, p. 71, 2016.

[71] M. Connor, G. Canal, and C. Rozell, "Classification and pathology of lung cancer," *Surgical Oncology Clinics*, vol. 25, pp. 447–468, 2016.

[72] S. Norouzi, D. Fleet, and M. Norouzi, "Exemplar vae: Linking generative models, nearest neighbor retrieval, and data augmentation," *34th Conference on Neural Information Processing Systems (NeurIPS 2018)*, vol. -, 2020.

[73] J. Yoo, H. Lee, and N. wak, "Density estimation and incremental learning of latent vector for generative autoencoders," *Mathematics, Computer Science*, vol. -, 2019.

[74] L. Ternes, M. Dane, M. Labrie, G. Mills, J. Gray, L. Heiser, and Y. Chang, "Me-vae: Multi-encoder variational autoencoder for controlling multiple transformational features in single cell image analysis," *bioRxiv*, vol. -, pp. –, 2021.

[75] N. Simidjievski, C. Bodnar, I. Tariq, P. Scherer, A. T., and et. al., "Variational autoencoders for cancer data integration: Design principles and computational practice," *Frontiers in Genetics*, vol. 10, p. 1205, 2019.

[76] N. C. Institute, "The cancer genome atlas program," 2022. Last accessed: 10 November 2022.

[77] J. Klys, J. Snell, and R. Zemel, "Learning latent subspaces in variational autoencoders," *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, vol. -, 2018.

[78] M. Connor, G. Canal, and C. Rozell, "Variational autoencoder with learned latent structure," *24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. -, 2021.

[79] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Statist*, vol. 33, pp. 1065–1076, 1962.

[80] W. C. of Tumours, *Thoracic Tumours*, vol. 5. Lyon (France): International Agency for Research on Cancer: Editorial Board, 2021.

[81] D. L. Rimm, G. Han, and et. al., "A prospective, multi-institutional, pathologist-based assessment of 4 immunohistochemistry assays for pd-l1 expression in non-small cell lung cancer," *JAMA oncology*, vol. 3(8), p. 1051–1058, 2017.

[82] L. McInnes and et. al., "Umap: Uniform manifold approximation and projection for dimension reduction," 2018.

[83] D. Blei, A. Kucukelbir, and J. McAuliffe, "Variational inference: A review for statisticians.," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

A ELBO: Evidence Lower Bound

A.1 Standard ELBO

In Bayesian inference, we are often interested in the posterior distribution $p(z|x)$ where x are the observations, and z are latent variables. This evidence is hard to compute because we have introduced latent variables that must now be marginalized out. Such integrals are often intractable in the sense that, we do not have an analytic expression for them or they are computationally intractable [83].

The main challenge with the variational inference objective is that it implicitly depends on the evidence, $p(x)$. Because we cannot compute the desired KL divergence, we optimize a different objective that is equivalent to this KL divergence up to constant. This new objective is called the evidence lower bound or ELBO [83].

$$
\begin{aligned}
&\ln p(x) \\
&\quad = \ln \int_z p(x,z)\,dz \\
&\quad = \ln \int_z p(x,z)\frac{q(z|x)}{q(z|x)}\,dz \\
&\qquad \text{Jenssen's inequality: } \varphi(\mathrm{E}\{X\}) \le \mathrm{E}\{\varphi(X)\} \\
&\quad \ge \mathrm{E}_{q(z|x)}[\ln \frac{p(x,z)}{q(z|x)}] \text{ by Jenssen's inequality} \\
&\quad = \mathrm{E}_{q(z|x)}[\ln \frac{p(x|z)p(z)}{q(z|x)}] \\
&\quad = \mathrm{E}_{q(z|x)}[\ln p(x|z)] + \mathrm{E}_{q(z|x)}[\ln \frac{p(z)}{q(z|x)}] \\
&\quad = \mathrm{E}_{q(z|x)}[\ln p(x|z)] + \int_z q(z|x)\ln \frac{p(z)}{q(z|x)}\,dz \\
&\quad = \mathrm{E}_{q(z|x)}[\ln p(x|z)] + \int_z q(z|x)\ln p(z)\,dz - \int_z q(z|x)\ln q(z|x)\,dz
\end{aligned}
$$

$$
\begin{aligned}
&= \mathrm{E}_{q(z|x)}[\ln p(x|z)] + \int_z q(z|x) \ln \frac{p(z)q(z)}{q(z)}\, dz - \int_z q(x|z)q(z) \ln(q(x|z)q(z))\, dz \\
&= \underbrace{\mathrm{E}_{q(z|x)}[\ln p(x|z)]}_{\text{Average reconstruction}} + \underbrace{D_{KL}[q(z)||p(z)]}_{\text{KL between p and q}} \\
&= likelihood - KL
\end{aligned}
$$

A.1.1 Reconstruction term

$$
\text{Reconstruction} = \sum_{i=1}^{D} x_i \log \hat{x}_i + (1 - x_i) \log(1 - \hat{x}_i) \tag{A-1}
$$

This term is the binary cross entropy between the output $\hat{x}$ and the input x. This part maximizes the log-likelihood, the likelihood tries to make the generated image more correlated to the latent variable, which makes the model more deterministic.

A.1.2 KL term

This term minimizes the KL divergence between the posterior and the prior. We usually assume the prior as a standard Gaussian distribution, and minimizing the KL will make the posterior more similar to the prior, which means we are trying to make the posterior to be a smooth Gaussian distribution.

The following is the demonstration of the origin of the KL term with a Gaussian assumption, where $p(x)$ and $q(x)$ are probability distributions, μ is the mean and σ is the variance.

$$
p(x) = \mathrm{N}(\mu_1, \sigma_1) \;\text{ and }\; q(x) = \mathrm{N}(\mu_2, \sigma_2)
$$

$$
\begin{aligned}
&\mathrm{KL}(p, q) \\
&= \underbrace{-\int p(x) \ln q(x)\, dx + \int p(x) \ln p(x)\, dx}_{\text{Relative entropy}} \\
&\quad \text{from Bishop's Pattern Recognition and Machine Learning} \\
&= -\int p(x) \ln \frac{1}{(2\pi\sigma_2^2)^{(\frac{1}{2})}} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}\, dx + \int p(x) \ln p(x)\, dx \\
&= -\int p(x) \ln e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}\, dx + \int p(x) \ln(2\pi\sigma_2^2)^{(\frac{1}{2})}\, dx + \int p(x) \ln p(x)\, dx \\
&= -\int p(x)(-\frac{(x-\mu_2)^2}{2\sigma_2^2})\, dx + \frac{1}{2} \ln(2\pi\sigma_2^2) + \int p(x) \ln p(x)\, dx
\end{aligned}
$$

$$
\begin{aligned}
&= -\int p(x)(-\frac{(x^2-2x\mu_2+\mu_2^2)}{2\sigma_2^2})\,dx + \frac{1}{2}\ln(2\pi\sigma_2^2) + \int p(x)\ln p(x)\,dx \\
&= \frac{\int p(x)x^2\,dx - \int p(x)2x\mu_2\,dx + \int p(x)\mu_2^2\,dx}{2\sigma_2^2} + \frac{1}{2}\ln(2\pi\sigma_2^2) \\
&+ \int p(x)\ln p(x)\,dx \\
&= \frac{\mathrm{E}[x^2]-2\mathrm{E}[x]\mu_2+\mu_2^2}{2\sigma_2^2} + \frac{1}{2}\ln(2\pi\sigma_2^2) + \int p(x)\ln p(x)\,dx \\
&\quad var(x) = \mathrm{E}[x^2] - \mathrm{E}[x]^2 \text{ , then } \mathrm{E}[x^2] = \sigma_1^2 + \mu_1^2 \\
&= \frac{\sigma_1^2+\mu_1^2-2\mu_1\mu_2+\mu_2^2}{2\sigma_2^2} + \frac{1}{2}\ln(2\pi\sigma_2^2) + \int p(x)\ln p(x)\,dx \\
&= \frac{\sigma_1^2+(\mu_1-\mu_2)^2}{2\sigma_2^2} + \frac{1}{2}\ln(2\pi\sigma_2^2) + \int p(x)\ln p(x)\,dx \\
&= \frac{\sigma_1^2+(\mu_1-\mu_2)^2}{2\sigma_2^2} + \frac{1}{2}\ln(2\pi\sigma_2^2) - \frac{1}{2}(1+\ln(2\pi\sigma_1^2)) \\
&= \frac{\sigma_1^2+(\mu_1-\mu_2)^2}{2\sigma_2^2} + \frac{1}{2}\ln(2\pi\sigma_2^2) - \frac{1}{2}\ln(2\pi\sigma_1^2) - \frac{1}{2} \\
&= \frac{\sigma_1^2+(\mu_1-\mu_2)^2}{2\sigma_2^2} + \frac{1}{2}\ln(\frac{2\pi\sigma_2^2}{2\pi\sigma_1^2}) - \frac{1}{2} \\
&= \frac{\sigma_1^2+(\mu_1-\mu_2)^2}{2\sigma_2^2} + \ln(\frac{\sigma_2^2}{\sigma_1^2})^{\frac{1}{2}} - \frac{1}{2} \\
&= \frac{\sigma_1^2+(\mu_1-\mu_2)^2}{2\sigma_2^2} + \ln(\frac{\sigma^2}{\sigma_1}) - \frac{1}{2} \\
&\quad \mathrm{KL}(p,q) = 0 \text{ when } \mu_1 = \mu_2 \text{ and } \sigma_1 = \sigma_2
\end{aligned}
$$

ELBO

$$
\begin{aligned}
&= -\frac{1}{2}\sum_{j=1}^{J}(1+(\sigma_j^2)-(\mu_j^2)-(e^{\sigma_j^2})) \\
&+ \sum_{i=1}^{D} x_i \log\hat{x}_i + (1-x_i)\log(1-\hat{x}_i)
\end{aligned}
$$

A.2 Regularized ELBO

Higgings et.al. (2017) [66] explore an approach that introduces an adjustable hyperparameter β that modulates the learning constraints applied to the model, while balancing the trade-off between the model capability to reconstruct the input image and the understanding of representation space. If the value of the parameter $\beta = 1$, corresponds to the base model of

the VAE [67], with $\beta > 1$, the model is pushed to learn a more efficient latent representation for the data, untangling the generated space. On the contrary, if $\beta < 1$ the model tries to reconstruct the input data in a better way.

ELBO

$$\underbrace{\mathrm{E}_{q(z|x)}[\ln p(x|z)]}_{\text{Average reconstruction}} + \underbrace{D_{KL}[q(z)||p(z)]}_{\text{KL between p and q}}$$

$$= \underbrace{\sum_{i=1}^{D} x_i \log \hat{x}_i + (1 - x_i) \log(1 - \hat{x}_i)}_{\text{Reconstruction term}} - \underbrace{\frac{\beta}{2J} \sum_{j=1}^{J} (1 + (\sigma_j^2) - (\mu_j^2) - (e^{\sigma_j^2}))}_{\text{KL term}}$$