

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

# Deep Learning Solutions for Lung Cancer Characterization in Histopathological Images

João Moranguinho Bastardo Moura



Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Tânia Pereira

Co-Supervisor: Hélder Oliveira

Co-Supervisor: José Luis Costa

February 17, 2021



# **Deep Learning Solutions for Lung Cancer Characterization in Histopathological Images**

**João Moranguinho Bastardo Moura**

Mestrado Integrado em Engenharia Informática e Computação

February 17, 2021



# Abstract

Cancer is one of the leading death causes in the world, specifically, lung cancer. According to the World Health Organization, at the end of 2020, around 2.2 million people were diagnosed with lung cancer, and 1.8 million fatalities resulted from it. Correctly identifying its presence in a patient and classifying its subtype and stage is fundamental for the appropriate target therapy adoption.

One of the gold standards used to identify and classify cancer is the microscopic visual inspection of histopathological images *i.e.* small tissue samples excised from a patient. Expert pathologists are responsible for this inspection. However, it requires a significant amount of time and sometimes leads to non-consensual results.

With the growth of computational power and data availability, modern Artificial Intelligence solutions can be developed to automate and speed up this process. Deep Neural Networks using histopathological images as an input currently embodies state-of-the-art in automated lung cancer diagnostic solutions, with Convolutional Neural Networks achieving the most compelling accuracies in tissue type classification. One of the main reasons for such results is the increasing availability of voluminous amounts of data, acquired through the efforts employed by extensive projects like The Cancer Genome Atlas.

Nonetheless, histopathological images remain weakly labelled/annotated. Most common pathologist annotations refer to the entirety of the image and not to individual regions of interest in the patient's tissue sample. Recent works have demonstrated Multiple Instance Learning as a successful approach in classification tasks entangled with this lack of annotation, by representing images as a bag of instances where a single label is available for the whole bag.

Thus, we propose a bag/embedding-level lung tissue type and subtype classifier using a Convolutional Neural Network in a Multiple Instance Learning approach, where the automated inspection of lung histopathological images determines the presence of cancer, and its possible subtype, in a given patient. Furthermore, we employ a post-model interpretability algorithm to validate our model's predictions and highlight the regions of interest for such predictions.

**Keywords:** Artificial Intelligence, Deep Learning, Classification and Prediction, Lung Cancer, Histopathology, Genetic Mutations



# Resumo

Nos dias de hoje, o cancro destaca-se como sendo uma das doenças responsáveis pelo maior número de óbitos. Em particular o cancro do pulmão. De acordo com a Organização Mundial de Saúde, no fim do ano de 2020, esta doença foi diagnosticada a cerca de 2.2 milhões de pessoas e levou a 1.8 milhões de óbitos. O seu diagnóstico correto e atempado é extremamente crucial, assim como a identificação do seu subtipo e estado de desenvolvimento. O sucesso deste diagnóstico permite a adoção de terapias específicas a cada paciente, aumentando a sua probabilidade de sobrevivência.

Uma das ferramentas mais comuns no diagnóstico de cancro é a inspeção microscópica de imagens histopatológicas. Estas imagens são pequenas amostras de tecido removido de pacientes portadores de tumores suspeitos. A inspeção destas imagens não é simples e requer o esforço de patologistas com formação muito específica. Os patologistas responsáveis por esta análise tendem a inspecionar um número elevado de amostras, com grande variabilidade entre elas. Este processo torna-se então exaustivo e pode levar a diagnósticos pouco consensuais entre colegas.

Com o crescimento do poder computacional e da disponibilidade de dados, surgem soluções modernas baseadas em Inteligência Artificial para automatizar e acelerar este processo. O estado-da-arte desta automatização é dominado por Redes Neurais Artificiais constituídas por múltiplas camadas. Um especial destaque é atribuído às Redes Neurais Convolucionais, capazes de atingir os níveis de precisão mais elevados na literatura. Uma das grandes razões para o sucesso destas soluções é a disponibilidade de enormes quantidades de dados, abertos ao público, gerados através dos esforços de projetos como o The Cancer Genome Atlas.

Apesar de existirem estes enormes volumes de dados, as anotações dos mesmos e a sua categorização mantêm-se um pouco limitadas. No caso das imagens histopatológicas, as anotações dos patologistas indicam maioritariamente o diagnóstico atribuído à amostra de tecido como um todo. Contudo, não fazem referência às pequenas regiões das imagens que têm maior impacto no diagnóstico. Multiple Instance Learning é uma corrente de investigação focada em contornar esta limitação. Nesta corrente, as imagens originais são consideradas como quadro composto por diversas peças. Esta formulação permite que as Redes Neurais consigam utilizar as amostras sem conhecer quais as regiões mais impactantes e mesmo assim sejam capazes de as prever.

Assim, propomos um classificador de imagens histopatológicas do pulmão, capaz de detetar a presença de cancro, o seu subtipo e possíveis mutações genéticas, utilizando Redes Neurais Convolucionais num contexto de Multiple Instance Learning. Adicionalmente, integramos um algoritmo de interpretabilidade nos nossos modelos, com o objetivo de validar as suas capacidades de classificação. Esta validação é feita através marcação das regiões das imagens que são consideradas mais importantes para a sua classificação.

**Palavras Chave:** Inteligência Artificial, Deep Learning, Classificação e Previsão, Cancro do Pulmão, Histopatologia, Mutações Genéticas





# Acknowledgements

I want to start by mentioning the incredible opportunity provided by my supervisors, who allowed me to explore fields that I have been admiring for a long time now and wish to grow in. I would also like to mention the continuous, timely and caring support I have received from them. An enormous appreciation goes out to you. Another mention must be attributed to those who have been by my side throughout all of the years leading up to this moment. Family, friends and acquaintances that helped me grow and develop my personality and love for the things I do. To all of you, I give out a long-lasting and robust hug, hoping that the love you gave me will be reflected in my work and lifetime accomplishments.

On behalf of my supervision, colleagues and myself, I also want to acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health for the free publicly available "The Cancer Genome Atlas (TCGA)" Database used in this work.

João Moranguinho Bastardo Moura



*“The game of science is, in principle, without end.  
He who decides one day that scientific statements do not call for any further test,  
and that they can be regarded as finally verified, retires from the game.”*

Karl R. Popper



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	2
1.2	Objectives . . . . .	2
1.3	Contributions . . . . .	2
1.4	Document Structure . . . . .	3
<b>2</b>	<b>Background</b>	<b>5</b>
2.1	Histopathology and Digital Histopathology . . . . .	5
2.2	Normal and Cancer Cell Features in Histopathological Images . . . . .	6
2.3	Lung Cancer Types and Subtypes . . . . .	6
2.4	Genetic Mutations . . . . .	8
2.4.1	Squamous Cell Carcinoma . . . . .	8
2.4.2	Adenocarcinoma . . . . .	9
2.5	Artificial Neural Networks and Deep Learning . . . . .	10
2.6	Convolutional Neural Networks . . . . .	12
2.7	Multiple Instance Learning Applied to Neural Networks . . . . .	14
2.8	Explainability and Interpretability in Machine Learning . . . . .	15
2.9	Summary . . . . .	16
<b>3</b>	<b>Literature Review</b>	<b>17</b>
3.1	Cancer characterization models using histopathologic images . . . . .	17
3.1.1	Machine Learning . . . . .	17
3.1.2	Deep Learning . . . . .	18
3.2	Genetic Mutations identified in histopathologic images . . . . .	21
3.3	Summary . . . . .	22
<b>4</b>	<b>Data Sets Characterization</b>	<b>25</b>
4.1	Histopathological Images . . . . .	25
4.2	Genetic Mutation Data . . . . .	28
4.3	Summary . . . . .	29
<b>5</b>	<b>Lung Cancer Characterization in Histopathological Images</b>	<b>31</b>
5.1	Tissue Type Classification Model . . . . .	32
5.1.1	Data Gathering and Categorisation . . . . .	32
5.1.2	Data Preprocessing . . . . .	33
5.1.3	Baseline Classifier . . . . .	34
5.1.4	Attention-Based Classifier . . . . .	35
5.1.5	Training, Validation and Testing . . . . .	37

5.1.6	Results . . . . .	37
5.1.7	Attention Maps . . . . .	38
5.1.8	Discussion . . . . .	39
5.2	Lung Cancer Subtype Classification Model . . . . .	40
5.2.1	Data Gathering and Categorisation . . . . .	40
5.2.2	Data Preprocessing . . . . .	41
5.2.3	Attention-Based Classifier . . . . .	42
5.2.4	Training, Validation and Testing . . . . .	42
5.2.5	Results . . . . .	42
5.2.6	Attention Maps . . . . .	42
5.2.7	Discussion . . . . .	43
5.3	Lung Cancer Gene Mutation Classification Model . . . . .	44
5.3.1	Data Gathering and Categorisation . . . . .	44
5.3.2	Data Preprocessing . . . . .	45
5.3.3	Attention-Based Classifier . . . . .	46
5.3.4	Training, Validation and Testing . . . . .	46
5.3.5	Results . . . . .	46
5.3.6	Discussion . . . . .	47
5.4	Summary . . . . .	48
<b>6</b>	<b>Conclusions and Future Work</b>	<b>49</b>
	<b>References</b>	<b>51</b>

# List of Figures

2.1	Histopathological images from TCGA-LUAD dataset . . . . .	5
2.2	Summary of visual anomalies present in cancer cells . . . . .	6
2.3	Lung cancer locations summarized by types and subtypes . . . . .	7
2.4	Visual clues for the presence of squamous cell carcinoma . . . . .	7
2.5	Visual clues for the presence of adenocarcinoma . . . . .	8
2.6	Commonly Mutated Genes in squamous cell carcinomas of the lung . . . . .	9
2.7	Commonly Mutated Genes in lung adenocarcinoma patients . . . . .	10
2.8	Artificial Neuron and Artificial Neural Network architecture . . . . .	12
2.9	Convolution operation . . . . .	12
2.10	Convolution operation over a MNIST data sample . . . . .	13
2.11	Hierarchical relationship between patterns . . . . .	13
2.12	Traditional supervised learning and Multiple Instance Learning . . . . .	14
2.13	Introduction of prediction interpretability with GradCam algorithm . . . . .	16
3.1	CNN for lung nodule detection in chest radiographs . . . . .	18
3.2	Ensemble strategy for colorectal cancer classification . . . . .	19
3.3	CNN for skin cancer classification . . . . .	20
4.1	Histopathological images from TCGA-LUAD dataset . . . . .	27
4.2	Histopathological images from TCGA-LUSC dataset . . . . .	27
4.3	Distribution of image width and height for TCGA datasets . . . . .	28
4.4	Distribution of mutated genes by number of cases. . . . .	28
5.1	40x and 5x magnification comparison . . . . .	31
5.2	TCGA-Barcode file name format . . . . .	32
5.3	Folder structure used to store tissue type classification data . . . . .	33
5.4	Image tilling and scoring process . . . . .	34
5.5	Summary of classical pooling techniques and novelty one . . . . .	36
5.6	Tissue Type Classification Model Pipeline . . . . .	37
5.7	Attention Maps Generated with GradCam algorithm . . . . .	39
5.8	Folder structure used to store lung cancer subtype classification data . . . . .	41
5.9	Tilling at 5x magnification . . . . .	41
5.10	Keratin pearl highlight with GradCam . . . . .	43
5.11	Patients mutations dictionary . . . . .	45
5.12	Folder structure used to store genetic mutation classification data . . . . .	45





# List of Tables

4.1	TCGA-LUAD dataset characterization . . . . .	26
4.2	TCGA-LUSC dataset characterization . . . . .	27
5.1	Best tissue type classification results for each attention mechanism. . . . .	38
5.2	Tissue type classification performance comparison. . . . .	39
5.3	Best cancer subtype classification results for each attention mechanism. . . . .	42
5.4	Lung cancer subtype classification performance comparison. . . . .	43
5.5	Results from TP53 gene mutation prediction in TCGA-LUAD set. . . . .	46
5.6	Results from KRAS gene mutation prediction in TCGA-LUAD set. . . . .	47
5.7	Genetic mutation classification performance comparison. . . . .	47



# Abbreviations

AI	Artificial Intelligence
AUC	Area Under the Curve
CNN	Convolutional Neural Network
GPU	Graphics Processing Unit
H&E	Hematoxylin and Eosin
MIL	Multiple Instance Learning
ML	Machine Learning
NSCLC	Non-Small Cell Lung Cancer
ROC	Receiver Operating Characteristic
SCLC	Small Cell Lung Cancer
TCGA	The Cancer Genome Atlas



# Chapter 1

## Introduction

Recent numbers published by the World Health Organization and American Cancer Society show that lung cancer was the second leading cancer diagnosed to patients in the year 2020. These numbers also point out that lung cancer was the leading cause for cancer-related fatalities registered this year [27]. Limiting lung cancer's impact starts with patients being correctly diagnosed in early stages, allowing the application of proper target therapies, thus avoiding possible unnecessary side effects associated with standard practices such as chemotherapy and radiotherapy. Most commonly reported side effects associated with lung cancer patients are toxicity, nausea, fatigue and shortness of breath [28].

One of the gold standards used for the classification and characterization of a patient's status is the microscopic inspection of histological slides. These slides are obtained by thinly slicing a pre-processed and thoroughly prepared paraffin-embedded sample of tissue excised from a patient's suspicious region. Pathologists look for many insights in these slides to aid them elaborating a diagnosis. These insights can arise from cell count, cell shape, nucleus size and shape, necrosis and many other features present in the tissue slide [54, 33]. However, the job of inspecting a significant and diverse amount of these slides can be quite tiresome and time-consuming, leading to potentially incorrect or non-consensual observations among colleagues [52]. Therefore the research in speeding up and automating this visual inspection retains tremendous interest.

Artificial Intelligence (AI) has proven its strengths and benefits in aiding professionals overcome issues present in situations like these by automating such tasks and processes, while demonstrating an above-average accuracy and precision, and sometimes revealing insights that escape the human eye. The automation of these processes is currently done through histopathological image tiles as an input for the training process of Convolutional Neural Networks (CNN).

## 1.1 Motivation

The medical imaging field has dramatically benefited from AI and Machine Learning (ML). Such advances include the development of Deep Learning solutions using CNN's for image classification tasks.

Research in the classification of lung histopathological images is not slim. Several tissue type classifiers using CNN architectures have been developed to identify normal and cancerous tissue. Additionally, further advances even lead to the development of subtype classifiers, capable of discerning the lung cancer subtype present in the tissue, *e.g.* squamous cell carcinoma and adenocarcinoma. However, most of these models function as "black box", and the rationale behind the decision making is not passed onto the end-user. Thus, there is an increased need for a pipeline capable of generating correct predictions and making them interpretable. Furthermore, a lack of in-depth research is still present in the identification of genetic mutations through the inspection of histopathological images. This lack of knowledge creates an excellent opportunity for novelty research as pathologists cannot identify such gene mutations by solely visually inspecting a tissue sample.

## 1.2 Objectives

Our works focal point is automating the visual inspection of lung histopathological images using modern Deep Learning solutions based on CNN's. This processes automation aims to speed up the diagnosis process, reducing the time patients wait for their doctors' response. Moreover, our work aims at increasing consensus between pathologists by creating a framework capable of generating predictions with high accuracy and interpretable insights.

Aware that lung cancer classification using histopathological images is not a novelty, our work envisions expanding such research by identifying relevant advances and current limitations. Moreover, a significant effort is concentrated in introducing interpretability into the model's predictions. The developments in this research area allow the model's performances validation and an increase in trust from pathologists.

The gap between modern classification approaches and identifying a patient's genetic mutational status (identified in the previous subsection of this document) is bridged in the developed application, informing pathologists of the mutation status of a set of genes, *i.e.* EGFR, TP53, KRAS. The latter aids pathologist in the assignment of precise diagnosis and appropriate forms of target therapy, thus increasing patients quality of life and survival.

## 1.3 Contributions

The contributions that surface from our work encapsulate three different classifiers, each one functioning in a specific hierarchical level, from tissue type to subtype to genetic mutational status.

These classifiers help pathologists visually inspect histopathological images by generating predictions and highlighting the regions responsible for such predictions. The following items present a succinct characterization of each level of classification in our hierarchy as well as an important feature ensured throughout them:

- **Tissue Type Classification Model** — A Deep Convolutional Neural Network model using a Multiple Instance Learning approach to distinguish normal and tumorous tissue in lung cancer patients;
- **Lung Cancer Subtype Classification Model** — A Deep Convolutional Neural Network model using a Multiple Instance Learning approach to identify the subtype of lung cancer present in histopathological images;
- **Lung Cancer Gene Mutation Classification Model** — A Deep Convolutional Neural Network model using a Multiple Instance Learning approach to identify the genetic mutations associated with lung cancer histopathological images;
- **Post-Model Interpretability Algorithm** — Interpretability algorithm to validate models performance and produce visual insights behind the predictions;

## 1.4 Document Structure

The present document is structured as follows. Chapter 2 presents a succinct description of the background and history associated with the concepts debated such as lung cancer origins, types and subtypes, histopathology and the visual inspection of histopathological images, neural networks and deep learning. Chapter 3 describes the state-of-the-art methodologies built to apply deep learning techniques to automate histopathological images visual inspection, along with the experimental results demonstrated by the authors. Chapter 4 lists and describes the datasets used for the training, validation and testing processes for the developed models. Chapter 5 gathers the proposed methodologies and workflows followed in developing our solutions. The last chapter, Chapter 6, brings to light the main insights gathered from the literature review process and the main lines of thought to guide future research and development.





## Chapter 2

# Background

To tackle the proposed challenge and to be able to develop the proposed solution, a set of concepts were reviewed and described in this section. This set of concepts reaches a broad spectrum of fields, ranging from pathology to computer science.

### 2.1 Histopathology and Digital Histopathology

Histopathology is the medical speciality referent to the study of pathologies through the visual inspection of tissue samples excised from patients' suspicious areas. To this day, this practice remains a gold standard in cancer diagnosis.

The patients considered here are referred for lung cancer, and thus, the primary locations of acquisition of the tissue samples are the top and bottom parts of the lungs. The collected samples follow a long and meticulous process of fixation, embedding, slicing and staining so that pathologists can accurately probe them under a microscope. Figure 2.1 shows an example of two images of processed excised tissue samples (histopathological images) from a patient in the TCGA-LUAD cohort, diagnosed with Non-Small Cell Lung Cancer subtype adenocarcinoma.

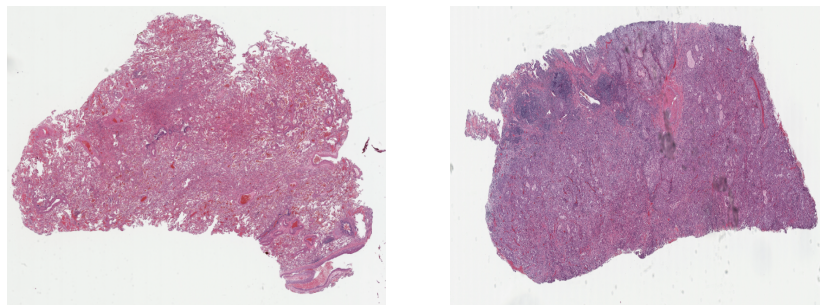


Figure 2.1: Histopathological images from TCGA-LUAD dataset [45].

Through the advances in hardware and technology, glass tissue slides can now be digitised and stored in databases, creating an opportunity for pathologists to rapidly share these slides with other colleagues generating more robust and consensual diagnostics. The largest and most commonly used publicly available databases, composed of hundreds of tissue slide specimens and their patient’s genomic information, were built from the many studies conducted by The Cancer Genome Atlas (TCGA) program. This program molecularly characterised primary cancer and normal samples spanning over the many known cancer types. This panoply of information led to many improvements in the field of oncology, including diagnosis, treatment and prevention.

## 2.2 Normal and Cancer Cell Features in Histopathological Images

When inspecting patients’ excised tissue samples under the microscope, pathologists look for several insights that can aid them in diagnosing pathologies. These insights may come from different levels of magnification/zoom. At the highest levels of magnification/zoom, pathologists can observe the cells composing the tissue sample. The visual features of these cells will determine the elaborated diagnosis. For instance, cancer cells exhibit anomalies when compared to normal cells [33]. Such anomalies include hyperchromatic nuclei and cytoplasm with variations in size and shape, multiple nuclei and smudged chromatin. Figure 2.2 summarises these visual anomalies.

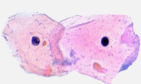
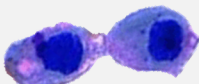
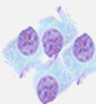
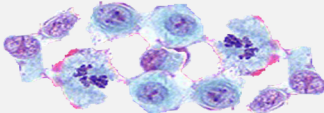

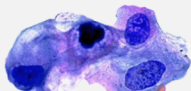
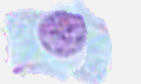
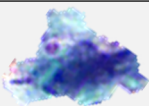
Normal	Cancer	
		Large, variably shaped nuclei
		Many dividing cells; Disorganized arrangement
		Variation in size and shape
		Loss of normal features

Figure 2.2: Summary of visual anomalies present in cancer cells, from LaMorte *et al.* [33].

## 2.3 Lung Cancer Types and Subtypes

Two main types of lung cancer have been identified and extensively studied to this day [54, 58]. The distinction between them originates in their location and physiological characteristics. Figure

2.3, adapted from Yoshinori *et al.* [1], summarises the most common lung cancer types, subtypes and their typical location in the lung.

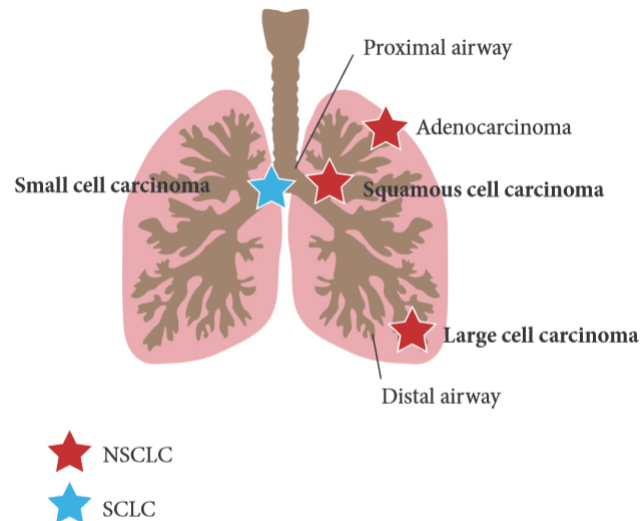


Figure 2.3: Lung cancer locations summarized by types and subtypes, from Yoshinori *et al.* [1].

The most common lung cancer type is Non-Small Cell Lung Cancer (NSCLC), and it is responsible for 80% to 85 % of the diagnosed cases. It encapsulates a set of subtypes that arise from different conditions but undergo similar treatment approaches and prognoses. This set is composed of the following:

- **Squamous Cell Carcinoma** — Originates in the squamous cells in the lung airways (bronchi and bronchioles) and is most commonly present in the central part of the lungs. Smokers account for 90% of the diagnosed squamous cell carcinomas. Visual clues of its presence include keratin pearl formation and keratin staining differentiation [54] (see Figure 2.4);

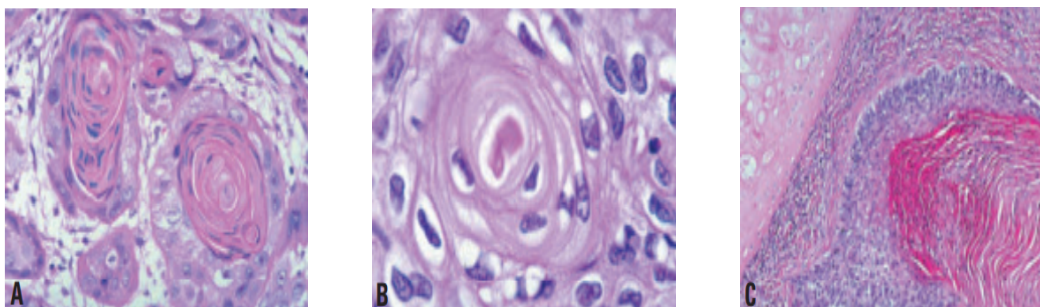


Figure 2.4: Visual clues for the presence of squamous cell carcinoma, from Travis *et al.* [54].

- **Adenocarcinoma** — Originates from the cells present in the lung alveoli and is most commonly present in peripheral zones of the lungs and pleura. Visual clues for its presence include glandular differentiation, mucin production and formation of acinar and papillary patterns [54] (see Figure 2.5);

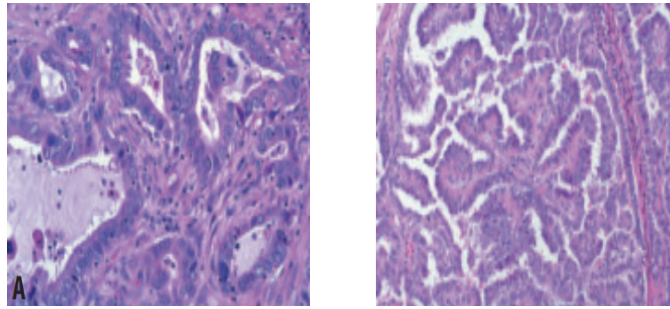


Figure 2.5: Visual clues for the presence of adenocarcinoma, from Travis *et al.* [54].

- **Large Cell Carcinoma** — Originates in epithelial cells in the lung. Also commonly present in the peripheral part of the lung.

The second most common lung cancer type, Small Cell Lung Cancer (SCLC), is responsible for the other 10% to 15% of diagnosis. Its fast growth and proliferation rate characterise this type, generally being metastasised by the first diagnosis. Furthermore, this type has a relatively positive response to therapy. However, it gravitates towards recurrence leading to reappearance in previously treated patients.

## 2.4 Genetic Mutations

Non inherited cancers mainly occur when key genes, responsible for cell growth and division and DNA repair, get mutated to a point where its functions are not correctly performed anymore, and unintended cancer cells can grow and proliferate uncontrollably, spreading throughout the original organ/tissue, or creating metastases in other distant parts of the body. Studies strive to identify the relationships between these gene mutations and the patient's cancer subtype, stage and prognosis. For example, the two most common NSCLC subtypes demonstrate significant differences in the most commonly mutated genes. Its proper identification can lead to a more robust diagnosis and application of therapeutic drugs, increasing patients' survival.

### 2.4.1 Squamous Cell Carcinoma

An extensive study conducted by The Cancer Genome Atlas Research Network [44] identifies TP53 as one of the most commonly mutated genes in squamous cell carcinomas, with almost all of the profiled cases exhibiting somatic mutations.

Somatic mutations of TP53 are prevalent among all types of cancer. This gene is responsible for producing a tumour suppressor protein (p53) that regulates cell growth and division. Hence, its mutation leads to an easier and faster proliferation of cancer cells as they grow and divide uncontrollably. A particular insight related with TP53 mutations in squamous cell carcinomas rises regarding prognosis. In other cancer subtypes, such as adenocarcinoma, somatic mutations of this gene represent a bad prognosis for the patient. However, in the case of squamous cell

carcinomas, the same does not apply [54]. Furthermore, a set of other commonly mutated genes was summarised in Figure 2.6. The percentages represented in the left-most part of the figure represent the percentage of cases exhibiting somatic mutations for each gene described in the right-most part of the figure.

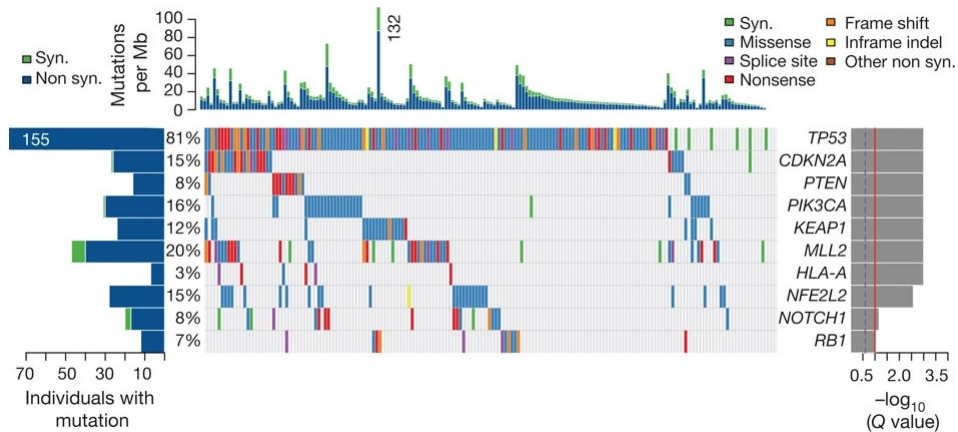


Figure 2.6: Commonly Mutated Genes in squamous cell carcinomas of the lung, from Cancer Genome Atlas Research Network [44].

One final crucial insight rose from this extensive study, as a set of candidate genes for therapeutic drug target were identified. These genes include PTEN, FGFR1, EGFR, PDGFRA, and KIT. With further advances in research, these genes may lead to target therapies able to replace chemotherapy.

### 2.4.2 Adenocarcinoma

In a similar fashion to that of squamous cell carcinoma, The Cancer Genome Atlas Research Network conducted an extensive study on the characterisation of lung adenocarcinoma [45]. In this study, TP53 gene mutations were also found to be the most common and were present in 46% of the profiled cases. As previously mentioned, in contrast to squamous cell carcinoma, adenocarcinoma patients enduring TP53 mutations have a poorer prognosis/survival [54].

Moreover, while relatively rare in squamous cell carcinoma cases, KRAS somatic mutations are commonly identified in 30% of adenocarcinoma patients. KRAS is also responsible for cell growth and division, which means that when mutated, cell growth and division become un-governed, leading to a favourable proliferation of cancer cells. One statistical insight related to the mutation of this specific gene is its higher frequency amongst smokers.

Additionally, most commonly mutated genes include KEAP1, STK11, EGFR and BRAF as displayed in Figure 2.7. The right side of plot (a) indicates the percentage of patients identified with somatic mutations for each gene in the left-most side of the plot. Plot (b) indicates the number of high and low transversion mutations. Transversion mutations refer to a specific type of point in which purine compounds (A's and G's) get switched by pyrimidine compounds (T's and C's). Finally, plot (c) presents some insights on the distribution of these mutated genes by gender.

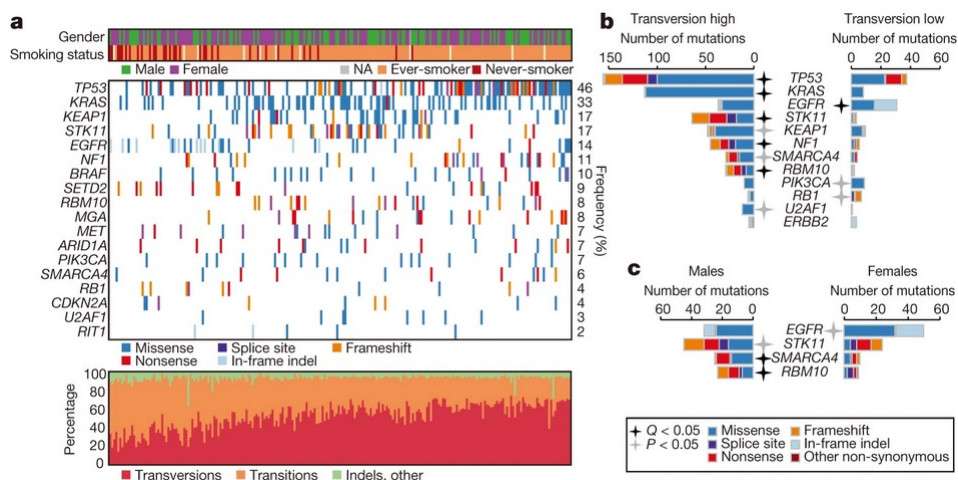


Figure 2.7: (a) Commonly Mutated Genes in lung adenocarcinoma patients; (b) Number of high and low transversion mutations for each gene; (c) Distribution of commonly mutated genes by gender, from Cancer Genome Atlas Research Network [45].

Contrary to the case of squamous cell carcinoma, this study does not summarise a set of candidate genes for therapeutic target. KRAS mutations remain primarily targeted with chemotherapy. Nonetheless, EGFR mutations have been extensively studied, and there are already target drugs ready to be administered in its treatment.

## 2.5 Artificial Neural Networks and Deep Learning

The rise of Artificial Neural Networks and Deep Learning applications is not recent. It is the product of a long history of work from scientists trying to define how the human brain operates and modelling its behaviours into computational systems.

The first characterisation of the human thought process had in its core the sequential associations we make when "thinking", theorising that thoughts were a specific list of sequential associations between concepts and objects that lead to a specific conclusion or action. This human cognition model is known as Associationism and was famously empirically demonstrated by Ivan Pavlov, in an experiment where the Conditional Reflex is formulated [41].

With the growth in scientific and technological advances, instruments like microscopes saw significant development and allowed the discovery of relevant insights on the human brain's composition. Its mass is mainly composed of an enormous number of interconnected neurons generating electrical signals amongst them. This composition is the foundation for the Connectionist models, introduced by Alexander Bain, in which the information is thought to be lodged in the connections between neurons and not in its cores. These theoretical models gave birth to connectionist machines, which paved the way for modern artificial neural networks.

Further in-depth inspection and research on the actual physiology of neurons extended these foundations, and in 1943 McCulloch and Pitts mathematically formulated the first model of a

logical-based neuron and later emulated boolean gates through the use of simple networks composed with this new logical-based neuron [42]. The model appeared promising and demonstrated relatively relevant capacities. However, it lacked a learning mechanism and could not infer new information.

Donald Hebb introduced a conceptual formulation of a learning method when he proposed the increase in neuron's weight/significance with its usage frequency. Simply put, increasing the rate at which a neuron A connects to a neuron B, will increase the neuron A ability to excite/activate neuron B. Even though not entirely in the same way, this mechanism remains one of the cornerstones of modern learning algorithms. Nonetheless, a significant flaw is inert to this model, which is the ever-growing instability caused by the increase in neurons weight without ever being possibly decreased, creating strength and belief in connections that were wrongfully formed.

The previously identified flaw was only tackled around 1958 when Rosenblatt unveiled the Perceptron, which was, as the media put it, the solution for everything [48]. The proposed model for the artificial neuron takes as an input the weighted sum of all the values from its connected neurons and excites/activates the next neuron if that sum surpasses a given threshold. Its ability to overcome the previously identified learning limitations comes from the updates conducted on the weights of the neurons regardless of the correctness of path. This means that correct paths increase the weights of neurons that lead to it, while incorrect paths decrease those same weights. This model displayed its strengths by easily mimicking the boolean gates experiment introduced by McCulloch and Pitts.

Still, a large wall remained unclimbed, as proven by Minsky and Papert [43]. The XOR, exclusive or, boolean gate was impossible to emulate with Rosenblatt's proposed model. With the help of a miss interpretation of Minsky and Papert's work, this constraint temporally stalled the belief and investment in the Artificial Neural Networks Field. The primary claim disregarded in this work was that the Perceptron model was not to be cancelled but used in a multi-layered structure, allowing it to excel in modelling the XOR gate and any other complex boolean function.

This multi-layered structure, where more than one layer of stacked artificial neurons is used instead of a single layer, is a pillar for the advances in Artificial Neural Networks and represents one of the earliest uses of Deep Learning. Figure 2.8, adapted from Kinsley *et al.* [31] displays the physiological inspiration for artificial neurons and the default architecture for Deep Learning Neural Networks.

The concept of Deep Learning is relatively intuitive as it refers to Artificial Neural Networks whose architecture is composed of more than two layers of stacked neurons. Several other layers may be used between the initial input layer and the final output layer. These layers, denoted as "hidden", turn the artificial neural network into a deep one as several levels of depth, *i.e.* several layers, are used to create predictions.

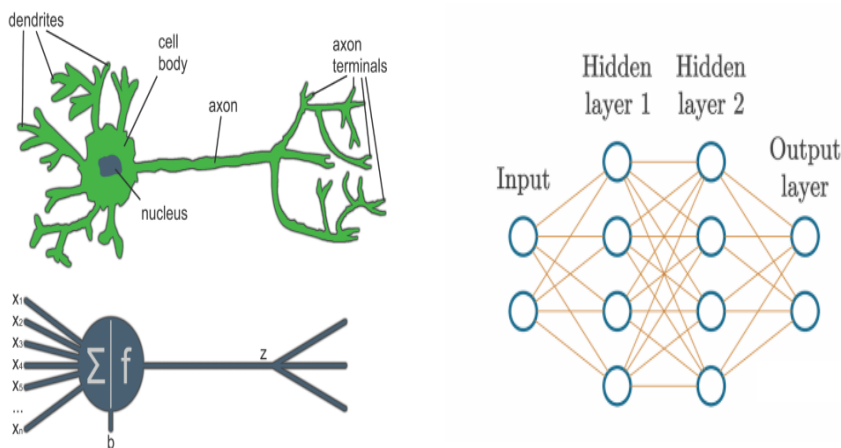


Figure 2.8: (a) Artificial Neuron inspired on the physiology of a human neuron; (b) Default architecture used in an Artificial Neural Network, from Kinsley *et al.* [31].

## 2.6 Convolutional Neural Networks

Convolutional neural networks are a type of deep artificial neural network characterised by converting simple input features into more complex outputs through the help of filters that convolute adjacent regions of interest. Figure 2.9, adapted from the "Deep Learning Book" by Ian Goodfellow *et al.* [18], demonstrates the structure of convolution operations.

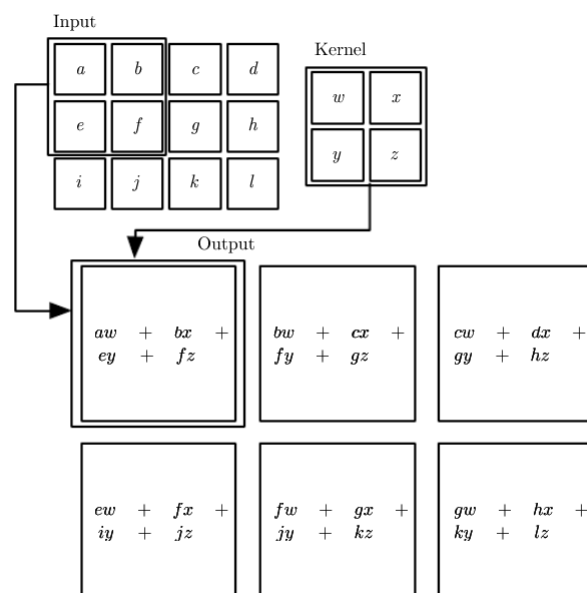


Figure 2.9: Convolution operation, from Goodfellow *et al.* [18].

In the example presented in this figure, the original 4x3 input image is convoluted with a 2x2 filter/kernel. At each step, the kernel window slides through different positions in the original image. With each step, a new cell in the output is generated by calculating the product between a



$2 \times 2$  region of the original input image and the  $2 \times 2$  filter. The resulting output is denominated as "response map", and it represents the response to the filter in different locations of the input. An example of this response map can be found in Figure 2.10, where the input image is a data sample from the MNIST dataset, and the generated result represents the locations of the original sample where the filter is most "salient".

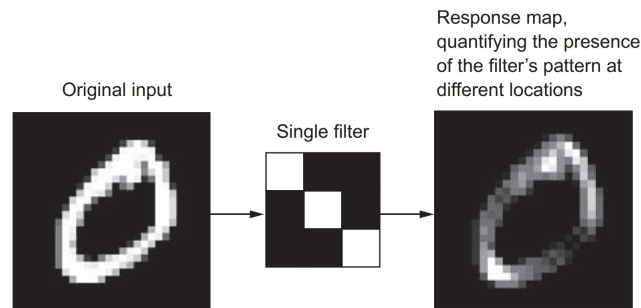


Figure 2.10: Convolution operation over a MNIST data sample, from Chollet *et al.* [9].

Convolutional Neural Networks are usually used in classification tasks where the input images remain relatively complex, and a single filter does not suffice. Thus, the convolutional layers of the model employ multiple filters and aggregate them in what is called a "feature map", a 3D space of amalgamated response maps generated for each filter. This is what allows models to pick up intricate patterns and form an idea of the relationships between them. This concept is particularly relevant as it introduces an idea of hierarchical structure between the different patterns identified in the images. Figure 2.11, adapted from the book "Deep Learning with Python" by François Chollet *et al.* [9], depicts an interesting example of these by highlighting the patterns that compose an image of a cat's face.

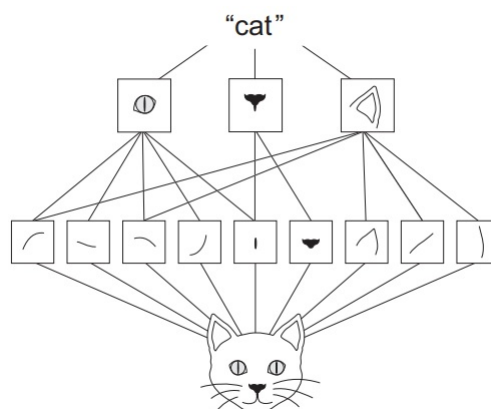


Figure 2.11: Hierarchical relationship between pattern identified in a cat's face, from Chollet *et al.* [9].

Initially, a set of simple patterns is identified. When amalgamated we understand they represent features the human-eye uses to identify a cat, such as the eyes, nose and ears. LeCun *et al.*

demonstrated one of the first significant applications of the potential of these neural networks in recognition of handwritten digits. This model was to be famously known as the LeNet [34].

The growing interest in this field and the application of such models led to competitions where researchers would prove their model’s novelty and accuracy. One of these competitions, ImageNet, took place in December 2012 and Krizhevsky along with his colleagues demonstrated a new winning model architecture named AlexNet [32]. The ImageNet challenge would be tackled with significant improvements again in 2014 when Simonyan and Zisserman developed a 19-layer model called OxfordNet. This model is commonly referred to as VGG19 [51]. The year of 2015 had two pivotal moments of significant improvement in the field of convolutional neural networks. On the one hand, the ImageNet challenge of that year was won by He et al. with a new model architecture called ResNet [21]. On the other hand, an impressive 22-layer network named GoogleNet was developed by Szegedy et al. [53], which benefitted from the use of inception-blocks, conceptualised by Lin and colleagues in 2013 [37].

Significant efforts from Google’s research teams have been moving Google’s Inception model forward. After its first debut, second and third versions have been developed and widely used. Improvements built on top of previous versions include batch normalisation. Google’s Inception has been the number one go-to in the field of medical imaging, as it remains highly performant under rigid memory constraints and highly complex problems.

## 2.7 Multiple Instance Learning Applied to Neural Networks

In ML tasks, supervision is one of the key steps for success and one of the main challenges. Models can be divided into three main categories regarding this concept, fully supervised, fully unsupervised and weakly supervised.

Multiple Instance Learning (MIL) is considered a supervised/weakly supervised ML technique where the data we wish to comprehend and make predictions on gets regarded as an amalgamation of smaller components [13, 6]. In the literature, data samples are denoted as **bags**, and its components are denoted **instances** (see Figure 2.12).

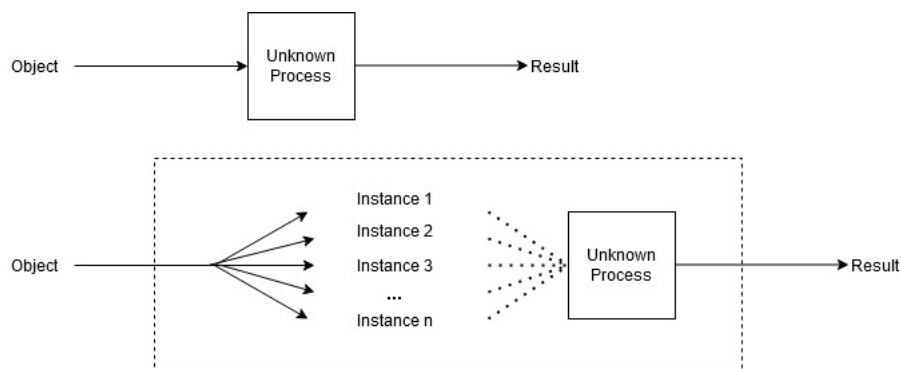


Figure 2.12: Traditional supervised learning and Multiple Instance Learning, adapted from Dieterich *et al.* [13].

In classification tasks using data sets comprised of images, a bag represents an entire image, and its instances represent some or all of its patches. This approach is particularly relevant in tasks where the classes of the data samples are known, while in contrast, the classes of its composing parts are not, *i.e.* a single label is available for the entirety of the bag, and no labels get provided for its instances.

MIL algorithms bifurcate into patch/instance-level classification and bag/embedding-level classification tasks. The former is concerned with the attribution of labels to the many individual instances present in each bag. The latter approach is only concerned with attributing a single label to each bag containing a set of instances, transforming our task into a fully-supervised classification one.

## 2.8 Explainability and Interpretability in Machine Learning

Explainability and Interpretability are two concepts that have begun to rise in popularity in the field of AI and ML. Researchers seek validation, transparency and trust for their models, envisioning the demystification of the classical black-box concern in which models produce predictions, but the end-user is unable to interpret its reasons and rationale. Achieving this goal can be done in one of two ways. First, by creating a model that in itself is understandable by humans and is capable of demonstrating the rationale behind its actions. This is the introduction of explainability on a model. Second, by introducing interpretability. In this case, the model itself does not have to be able to prove its rationale. Instead, it is concerned with generating outputs that can easily be understood/interpreted. To do so, interpretable models highlight the relationship between the inputs fed to the model and its predictions. In image classification tasks, using Deep Convolutional Neural Networks, this concept generally represents the highlight of specific regions of interest that the model uses as a foundation for its predictions.

A simple illustrative proof of concept is the animal image classification task using any CNN architecture. In this task, the model is concerned with predicting animal classes present in input images. Figure 2.13, adapted from Ramprasaath *et al.* [49] demonstrates an example of what the model's predictions could be, and the highlighted regions of interest that justify such predictions.

The introduction of this concept creates three possible levels of informational gain. The first one regards the justification for failure. In this case, models under-performing or over-performing in specific situations generate interpretable highlights that help the “user” understand what might be missing or introducing bias. The second level of informational gain brought by the introduction of this transparency in models is the visual validation of predictions, which leads to “users” increasing the trust they deposit in them. The third and final level is the most ambitious one and allows models that outperform humans in classification tasks to generate insights that escape the human eye and have the potential to become a new standard.

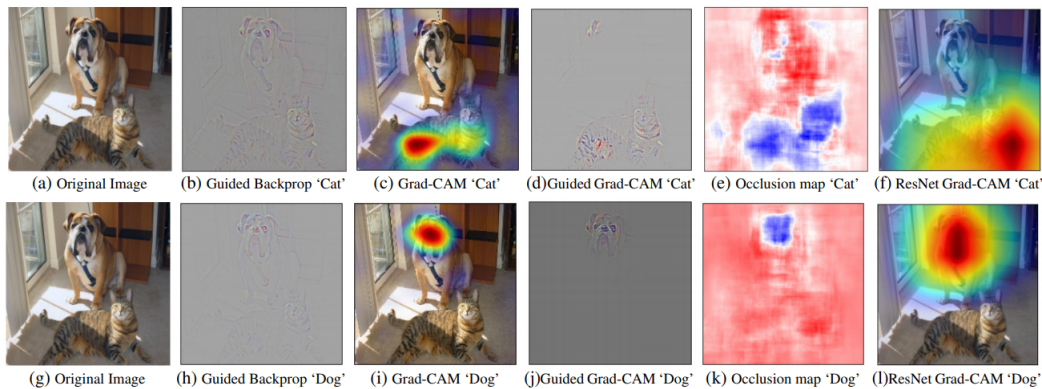


Figure 2.13: Introduction of prediction interpretability with GradCam algorithm, highlighting the regions significant to each class predicted by the model, from Ramprasaath *et al.* [49].

## 2.9 Summary

Visual inspection of histopathological images remains a gold-standard in lung cancer diagnosis. Expert pathologists inspect these images under a microscope and look for specific morphological features to give them insights related to the patients' diagnosis. In lung cancer patients and cancer patients in general, these features include cells with multiple, large and miss-shaped nuclei. Additionally, histopathological image inspection aids pathologists discerning the particular cancer subtype in a patient. However, this inspection does not rely solely on the previous cellular level morphological features. Each subtype of lung cancer must be probed for higher magnification level patterns. A great deal of interest also resides in identifying patients' genetic mutations through the inspection of these patterns.

Artificial Intelligence and Machine Learning algorithms, specifically Artificial Neural Networks, have demonstrated their potential in automating the inspection of histopathological images. In particular, Convolutional Neural Networks demonstrate the state-of-the-art accuracy in such task, by utilising various filters to identify the desired patterns in the original image and create a hierarchical relationship between them.

Nonetheless, one significant constraint remains present in these solutions. Most image annotations relate only to the whole tissue images cancerous status. Thus, no insights are provided on the specific regions exhibiting morphological characteristics justifying the predictions generated. Multiple Instance Learning approaches tackle this issue by decomposing images into smaller pieces and learning each one's significance. Additionally, interpretability algorithms such as Grad-Cam use the models' internal parameters to identify and highlight pieces exhibiting morphological features characterising predictions.

## Chapter 3

# Literature Review

This section presents the results obtained from a systematic literature review. This review process includes research on many fields of interest crucial to the development of our work. Moreover, these results encapsulate the accuracy and precision of traditional machine learning and modern Deep Learning techniques in classifying images gathered from suspicious tumours located in a set of organs or systems. Although lung cancer characterisation remains our research's primary interest, the innovation and breakthroughs achieved in classifying tissue from other organs and systems remain incredibly insightful.

### 3.1 Cancer characterization models using histopathologic images

#### 3.1.1 Machine Learning

Conventional Machine Learning techniques, such as Support Vector Machines (SVM's), Random Forests (RF's), Naive Bayes and K-Nearest Neighbors, use quantitative features to formulate predictions. These features commonly represent tissue characteristics such as nuclei and cytoplasm size, shape, and distance in histopathological image classification tasks. Beck *et al.* [4] conducted a study in which this set of features was used to train a prognosis model to effectively analyse breast histopathological images and generate patients 5-year survival prediction. However, this work is limited to the use of small image cores from tissue samples. Thus, a significant amount of information from the whole slide image is lost, and the model's ability to deal with heterogeneity is reduced. In addition to the previously mentioned features, additional quantitative information is present in the textural characteristics of the tissue samples like contrast and homogeneity. Sertel *et al.* [50] developed a multi-scale neuroblastoma classification model to identify stroma-rich and stroma-poor regions in histopathological images, based on these textural features. With the use of a multi-scale approach, the previously mentioned image size limitation is discarded, and the developed model reaches very compelling results. Nonetheless, a key limitation remained as no subtype differentiation insights were generated. One pivotal study tackling this issue was conducted by Yu

*et al.* in 2016 [57], where an extensive set of machine learning algorithms were employed and compared. These algorithms included SVM and RF to predict NSCLC subtypes in H&E images, acquired from TCGA and Stanford’s Tissue Microarray (TMA) publicly available datasets. To do so, these algorithms used a set of previously selected quantitative features to distinguish each class. The selection of said quantitative features was made with the aid of CellProfiler [7], which employed the measurement of cell shapes and sizes, pixel intensity and object’s textures. The results showed significant accuracy in subtype distinction, presenting an average AUC of 0.72 for the TCGA dataset and an average AUC of 0.78 for the TMA dataset.

Altogether, Machine Learning methods have demonstrated compelling results in histopathological image classification. However, they remain relatively scarce in the literature. The main reason for such scarcity is their heavy reliance on quantitative features. Modern research efforts demonstrate a more generous amount of interest in methods capable of identifying visual patterns and visual characteristics that can aid pathologists in their decision-making process, while also increasing the trust in the developed models. Such research interest emerges in the continuous development and extension of Deep Learning solutions.

### 3.1.2 Deep Learning

One of the earliest, and most significant, registers of medical imaging aid by Deep Learning technique comes from a lung nodule detection model for greyscale chest radiographs, developed by Lo *et al.* in 1995 [40]. This model was an initial version of CNN’s which performed two-dimensional convolutions on image regions. Figure 3.1 shows an example of the original chest radiograph images and the CNN architecture used. The results showed a significant increase in AUC when the novelty model was compared to its contemporary regular neural networks.

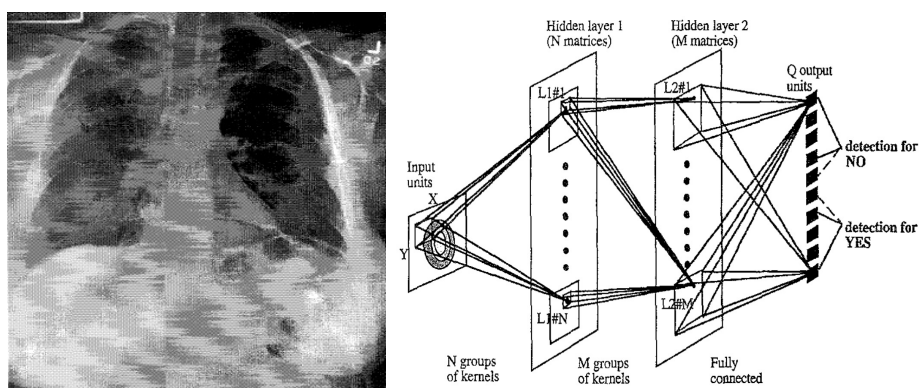


Figure 3.1: Chest radiograph and CNN architecture used in lung nodule detection, from Lo *et al.* [40].

Since then, AI and Deep Learning have become very popular in the medical imaging field, and many new opportunities and future challenges have been identified [19]. Extensive development has been applied in this field, across all specialities, as demonstrated by Litjens *et al.* [38] in their review of the extraordinary efforts moving Deep Learning algorithms and architectures forward.

The following subsections summarise some of the methods and findings surrounding the use of Deep Learning for cancer characterisation in histopathological images acquired from different organs (colon and rectum, skin, breast and lung).

### 3.1.2.1 Colorectal Cancer

Colorectal cancer is a grouped designation attributed to cancers originating in the colon or rectum, both included in the large intestine. At the end of 2020, it was responsible for 10% of the diagnosed cancers and around 10% of cancer-related fatalities. A Deep Learning solution for colorectal cancer characterization has been developed by Qaiser *et al.* [47], where they first combine a CNN trained with whole slide H&E stained images, with the patient's enhanced persistent homology profiles (ePHP). Later, significantly better results arise from a multi-stage ensemble learning strategy, combining the output of RF regression models attached to the ePHP and CNN (see Figure 3.3).

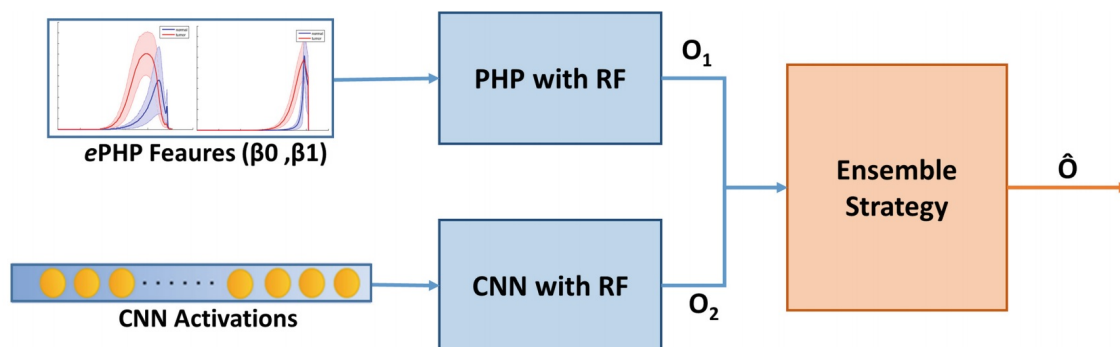


Figure 3.2: Pipeline for colorectal cancer classification, from Qaiser *et al.* [47].

Another study conducted by Bychkov *et al.* [5] shows that Deep Learning solutions can also predict patients survival. In this study, the authors employ a pre-trained VGG16 model to extract features from the original H&E stained images and subsequently feed these features into a Long Short Term Memory (LSTM) network. Such a method shows improvements in patients' disease-specific-survival prediction compared to the most common traditional machine learning classification algorithms, such as SVM, Logistic Regression and Naive Bayes. These results are in line with Qaiser *et al.* by demonstrating the potential for Deep Learning solutions applied to colorectal cancer when using histopathological images.

### 3.1.2.2 Breast Cancer

At the end of 2020, breast cancer was the most commonly diagnosed cancer, representing around 12% of all cases. Its staging and treatment decisions heavily rely on the detection of metastasis [2]. Yun Liu *et al.* [55] developed a metastases detection model using breast cancer histopathological images. This model improved the previous work by Wang *et al.* [39], which already demonstrated

significant results in detecting metastases. The new, improved model uses a CNN, specifically, Google's Inception V3 architecture, and its results show a close to pathologist-level accuracy, with an AUC value of 0.99 in the Camelyon16 data set. Recently, Xie *et al.* [56] extended breast cancer characterisation research by conducting a study in breast histopathological image classification. In this study, two CNN architectures were compared in the task of classifying cancerous and normal tissue images. The two chosen architectures were Inception ResNet V2 and Inception V3. The former presented the most compelling results with an AUC of 0.996, a value significantly higher than the previous works presented in the literature.

### 3.1.2.3 Skin cancer

Although not as common as the cancers mentioned above, skin cancer also holds a great deal of concern, and Deep Learning solutions can considerably aid its diagnosis. Esteva *et al.* [16] developed a CNN using the Inception V3 architecture to classify skin cancer patients through the inspection of photographic and dermoscopic images of skin lesions (see Figure 3.2). Experiments demonstrated the models high capacity for classification tasks, with a maximum AUC of 0.96 for both small and large data sets. These results show accuracy in classification close to that of professional pathologists, even when the data sets informational size is quite large. Another use of CNN's has outperformed eleven pathologists in the classification of histopathological melanoma images. The model developed by Heckler *et al.* [23] stood out in its testing phase, where it showed a mean sensitivity, specificity and accuracy higher than the ones exhibited by the eleven pathologists who classified the test data set.

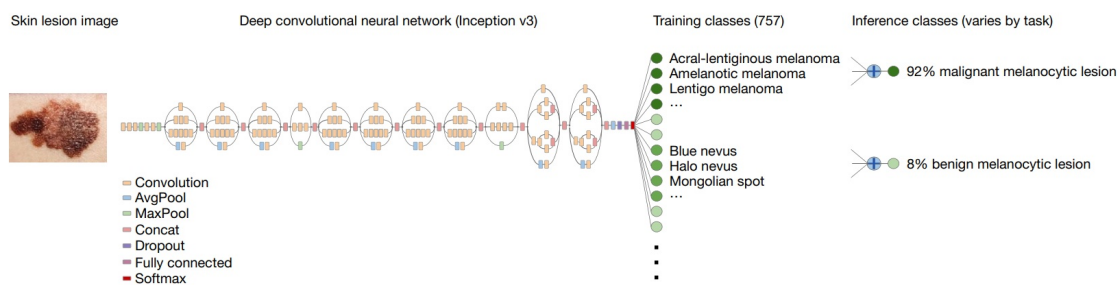


Figure 3.3: Pipeline for skin cancer classification using Google InceptionV3, from Esteva *et al.* [16].

### 3.1.2.4 Lung Cancer

The first significant concern in lung cancer characterisation is the detection of tumour tissue in the lung or other organs when it has metastasised. Several Deep Learning solutions have been developed to extend research on this concern. A recent study published by He *et al.* [22] proposed a framework for lung cancer whole slide image classification and annotation, where the first phase of the pipeline is characterised by a patch-based classification model using a CNN and the second



phase, the inference one, generates the annotations based on the heatmap created with the predicted probabilities for each patch. The proposed framework presented significantly compelling accuracy results and was capable of generating annotations similar to the ones produced by professional pathologists. As previously mentioned, one other way to identify the presence of lung cancer in a patient is through the inspection of metastases in other organs, specifically in lymph nodes. Another recent study published by Pham *et al.* [46] demonstrated significant results in lung cancer detection through the inspection of lymph node histopathological images. The proposed solution initially removed lymphoid follicles detected in the original images, which decreased and further aided the classification process. Overall an AUC of 0.922 was achieved in the detection of cancer.

The second principal concern is the identification of the cancer subtype, allowing pathologists to employ therapy solutions specific to each patient. In a set of studies aimed at correctly classifying NSCLC subtypes present in histopathological images, Hou *et al.* [24] employed what is now considered a state-of-the-art preprocessing technique for training data, where large images are split into same-sized tiles. These tiles trained the first level of the classification model, in which every patch was classified with a given probability of representing a subtype of NSCLC. This model's second level takes these probabilities and uses an SVM algorithm to label the whole image. With this structure, a compelling classification accuracy of 0.798 was achieved. Similarly, Khosravi *et al.* [29] developed an ensemble learning method capable of dealing with heterogeneous populations to classify different NSCLC subtypes and identify the biomarkers present in the tissue samples. To enhance the conducted studies informational gain, the authors underwent experiments with a range of CNN architectures for which the best results presented belong to Google's Inception V3. Additionally, Li *et al.* [35] conducted a study on thirty-three lung cancer patients where the whole slide images acquired from tumour tissue were classified into NSCLC subtypes. In this study, a set of CNN's was compared, when trained from scratch and fine-tuned. The results show that SqueezeNet [25], a relatively unknown smaller version of AlexNet that requires fewer parameters and storage size, holds up quite exceptional accuracy levels in classification. The authors demonstrate that the area under the ROC curve measure remains higher than the most common AlexNet, ResNet and VGG models, with values of 0.91 and 0.87 for the trained from scratch fine-tuned scenarios.

An analytical overview of these publications demonstrates the notable results obtained using Deep Learning solutions over conventional Machine Learning. Nonetheless, these publications lack to tackle research on the identification of genetic mutations, a concept that has gained a lot of momentum and represents great innovation value.

## 3.2 Genetic Mutations identified in histopathologic images

Apart from determining if a patient has cancer or not, and which subtype of cancer it has, histopathological images can help identify the patients' genetic mutations. This is critical to an appropriate diagnosis and treatment choice. Many studies point out a set of candidate genes, whose mutation commonly leads to precancerous lung lesions or invasive proliferation of advanced carcinomas.

This set of genes includes EGFR, KRAS, TP53 and STK11 as the most prevalent ones, although the incidence of some may be related to geographical and ethnic characteristics of the patients [8].

A recent study conducted by Nicolas Coudray *et al.* [10] shows unprecedented results on the classification of tissue whole slide images from the TCGA data sets regarding type and subtype, achieving pathologist-levels of accuracy. However, more importantly, the authors propose a deep CNN model, with a Google Inception V3 architecture, to detect genetic mutations commonly associated with the fast growth and development of lung cancers. Results show considerable accuracy in this detection process and indicate that further research can improve this task. The results also confirm the previously mentioned genes as the most common in lung cancer, as STK11, EGFR, FAT1, SETBP1, KRAS and TP53 were identified and accurately classified through the inspection of histopathological images. Another study published by Kim *et al.* [30] demonstrates the results of applying the previously mentioned deep CNN model to identify the genes mutated in histopathological images of melanomas (skin cancer). In this study, the two most commonly mutated genes in melanomas, BRAF and NRAS, were predicted with AUC's of 0.75 and 0.7, creating significant interest and future work opportunities. In a similar fashion Liao *et al.* [36] brings to light the strength of a less memory demanding deep CNN model identifying genetic mutations present in histopathological images from liver cancer patients. An average AUC of 0.7 was achieved in the prediction of a set of seven different genes. This set includes ALB, CSMD3, TNNB1, MUC4, OBSCN, TP53, and RYR.

The solutions presented in these publications are still unable to generate perfect predictions. However, the results obtained are reasonably promising, and further research may help achieve significant breakthroughs.

### 3.3 Summary

The literature review process gathers convincing evidence that Deep Learning solutions overperform conventional Machine Learning methods. Furthermore, CNN-based solutions are clearly denoted as the most common in medical imaging classification tasks. This review process also highlights the lack of research and development in models interpretability. Similarly, the tender state of genetic mutation classification has been reported.

A summarised view of the identified limitations present in the current state of the art for histopathological image classification is presented below:

- **Data Availability** — Publicly available cohort information containing histopathological images and genotype arrays remain slightly insufficient;
- **Data Storage** — Histopathological images tend to have high resolutions, thus, storing them along with all the complementary genotyping can be very memory demanding;
- **Class Imbalance** — Public datasets tend to be imbalanced, either containing a larger number of tumour samples or a larger number of healthy samples. Furthermore, the information

regarding mutated genes tends to be very sparse, increasing the difficulty in gathering a balanced set of classes for each gene;

- **Data Processing and Training Times** — Computational advances keep on increasing the speed at which machines process and use this data. However, it never seems fast enough, and with future improvements, greater speeds will be achieved.
- **Interpretability** — Most models presented in the literature lack an explainability/interpretability feature, as no insights on the model's "thought process" are generated. The only common tool used in this sense is the generation of heatmaps representing class prediction probabilities.



## Chapter 4

# Data Sets Characterization

A brief description of the publicly available datasets considered for the proposed work is provided in this section. These datasets encapsulate hundreds of histopathological images and genomic sequences acquired from lung cancer patients. Many of the state-of-the-art models for classification tasks use these images and sequences to gather benchmark results. The characteristics of the individual samples in the sets are also described including gender, the primary site of sample acquisition, disease type, patient’s vital status, the experimental strategy used for the acquired samples and its type.

### 4.1 Histopathological Images

The primarily identified datasets come from The Cancer Genome Atlas (TCGA), a research program born in 2005, aiming to molecularly characterize the genetic mutations present in cancer cells. This program runs under the National Cancer Institute’s Center for Cancer Genomics and the National Human Genome Research Institute which received its funding from the United States government.

The program extensively gathered data on cancer patients through genomic sequencing a tumour sample tissue excision. There is a vast amount of publicly available data on thirty-three different types of cancers [3], affecting different organs or systems, cataloguing the different identified mutations for each case and the resulting diagnosis. Ultimately this data has moved research forth and gave rise to many new improvements in the identification of new cancer target therapies, drugs and biomarkers. The following datasets belong to the extensive study in lung cancer characterization conducted by the TCGA program:

- **TCGA-LUAD** — Histopathological images and genomic sequences from lung adenocarcinoma patients from the TCGA-LUAD project cohort (see Table 4.1 and Figure 4.1);

- **TCGA-LUSC** — Histopathological images and genomic sequences from lung squamous cell cancer patients from the TCGA-LUSC project cohort (see Table 4.2 and Figure 4.2).

Each one of these datasets contains hundreds of Hematoxylin and Eosin (H&E) stained slides. This staining method allows the visual enhancement of different features in the tissue. Hematoxylin gives cell nuclei a blue colour, while eosin is responsible for staining cell cytoplasm with a pink colour. With this aid pathologists get a clear view of the cells in the tissue and address features that include size and shape, which are determinant for the diagnosis. The untrained eye might be able to partially distinguish between healthy, and tumour tissue images, as healthy tissue seems to be closer to a pink shade as nuclei remain relatively ordinary. Two examples of this distinction are exhibited in Figures 4.1 and 4.2. Nonetheless, the distinction between each subtype of cancer remains extremely difficult. Hence, the assessment of a trained pathologists remains critical.

Tables 4.1 and 4.2 summarize the information from the patients in each cohort, along with the experimental strategies used, sample type, and the identified commonly mutated genes. Moreover, Figures 4.3 plots the H&E image width and height distribution. TCGA-LUAD image samples have an average width of 50,000 pixels and an average height of 34,000 pixels. Similarly, TCGA-LUSC image samples have an average width of 45,000 pixels and an average height of 32,000 pixels.

Table 4.1: TCGA-LUAD dataset characterization

<b>Acronym</b>	<b>Description</b>
Cases	585
Gender	<i>Male(242) , Female(280)</i>
Primary Site	<i>Bronchus and lung</i>
Disease Type	<i>Adenoma and adenocarcinoma(549) , acinar cell neoplasms(22) Cystic/mucinous/serous neoplasms(11)</i>
Vital Status	<i>Alive(334) , Dead(188)</i>
Experimental Strategy	<i>WXS(582) , Methylation Array(579) , Genotyping Array(518) RNA-seq(515) , Tissue Slide(514) , miRNA-seq(513) Diagnostic Slide(478) , ATAC-seq(22)</i>
Sample Type	<i>Primary tumour(585), blood derived normal(424) Solid tissue normal(274), ffpe scrolls(2), recurrent tumour(2)</i>
Commonly Mutated Genes	<i>TP53, KRAS, EGFR, STK11, KEAP1</i>

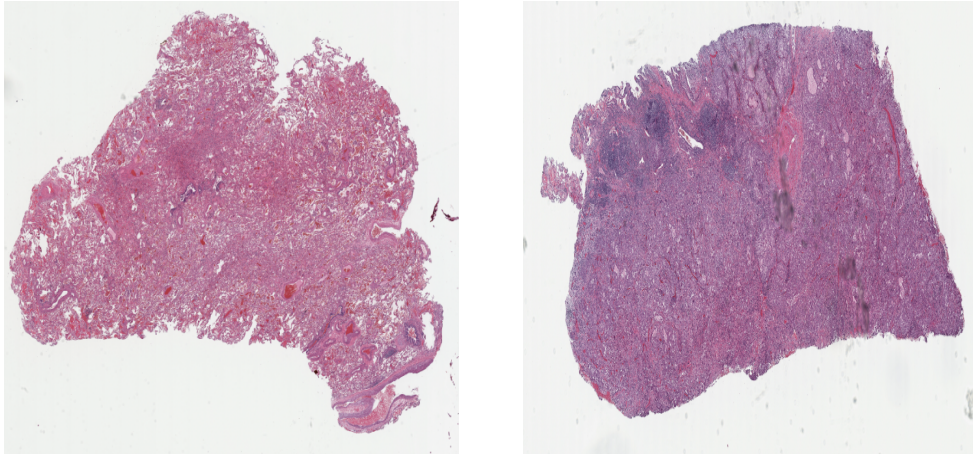


Figure 4.1: Histopathological images from TCGA-LUAD dataset

Table 4.2: TCGA-LUSC dataset characterization

<b>Acronym</b>	<b>Description</b>
Cases	504
Gender	Male(373) , Female(131)
Primary Site	Bronchus and lung
Disease Type	Squamous cell neoplasms(504)
Vital Status	Alive(284) , Dead(220)
Experimental Strategy	WXS(502) , Methylation Array(503) , Genotyping Array(504) RNA-seq(501) , Tissue Slide(495) , miRNA-seq(478) Diagnostic Slide(478) , ATAC-seq(16)
Sample Type	Primary tumour(504), blood derived normal(318) Solid tissue normal(254), ffpe scrolls(3)
Commonly Mutated Genes	TP53, MLL2, PIK3CA, CDKN2A, NF2L2, KEAP1

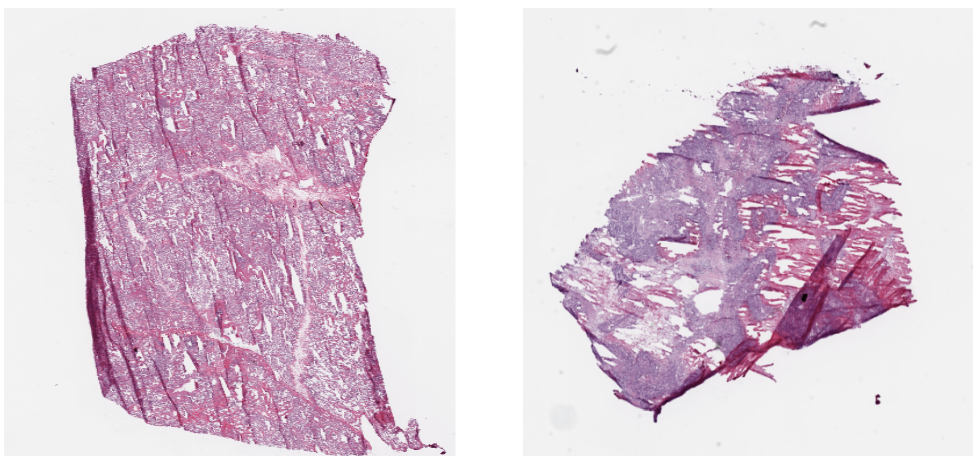


Figure 4.2: Histopathological images from TCGA-LUSC dataset

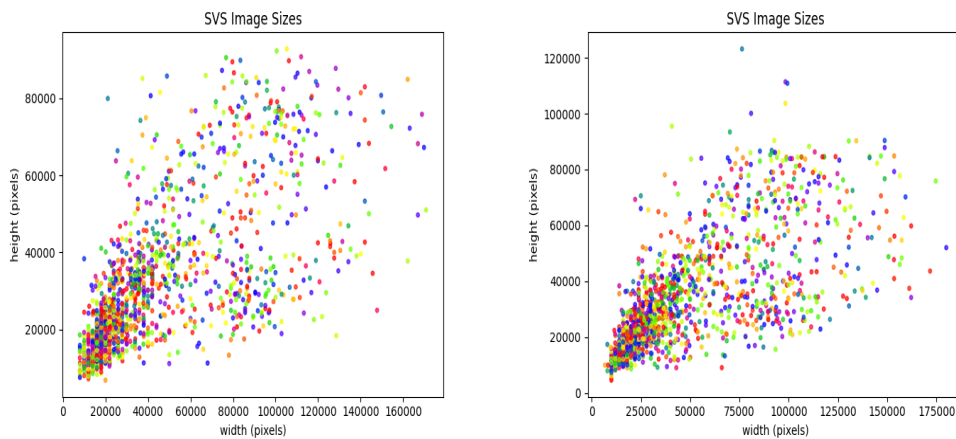


Figure 4.3: Distribution of image width and height for TCGA-LUSC and TCGA-LUAD datasets.

## 4.2 Genetic Mutation Data

Patient's whole genome sequence arrays compose another essential volume of information in TCGA datasets. This information allows pathologists to get a more in-depth insight into the actual subtype of cancer a patient has. From the original 585 patients belonging to TCGA-LUAD cohort, 448 (77%) exhibit at least one of the previously identified genetic mutations. On the other hand, out of the 504 patients belonging to TCGA-LUSC cohort, 442 (88%) of them also endure such genetic mutations. A slight increase in the percentage of patients with mutated genes is observed.

Figure 4.4 summarizes the number of patients associated with each mutation. The numbers comply with the statistics presented in Chapter 2.4. In the case of TCGA-LUAD patients, close to 30% of them endure KRAS mutations, while very few TCGA-LUSC do.

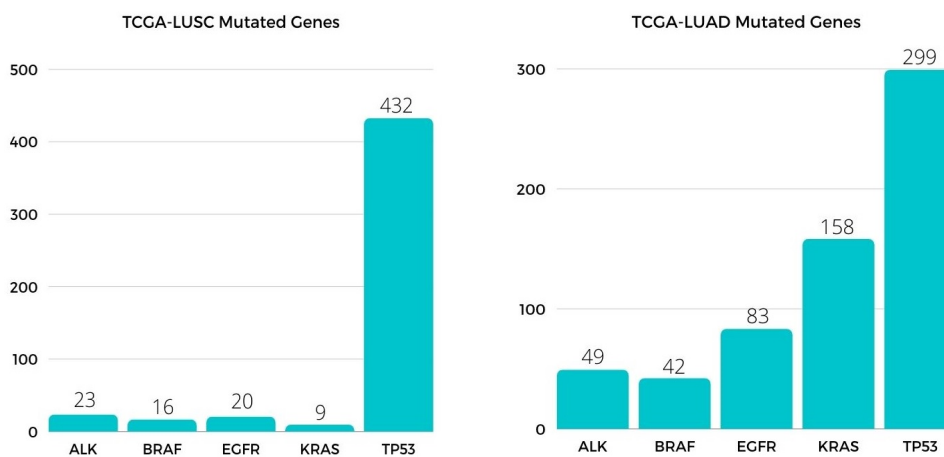


Figure 4.4: Distribution of mutated genes by number of cases.



### 4.3 Summary

The efforts conducted by research programs such as The Cancer Genome Atlas allow access to datasets crucial to our work. These datasets include hundreds of histopathological images, stained with H&E, which aids the distinction between normal and cancer samples, by giving the former a pink colouration and the latter a purple like colouration.

Additionally, these datasets also contain genomic sequencing arrays, which give us the patients' genetic mutation status. A statistical analysis of these data, allows us to formalize that squamous cell carcinoma patients commonly exhibit TP53 genetic mutations, while, on the other hand, adenocarcinoma patients tend to exhibit TP53, EGFR and KRAS mutations. This disparity in mutated genes, can further aid pathologists in elaborating a subtype diagnosis.



## Chapter 5

# Lung Cancer Characterization in Histopathological Images

This chapter gathers and describes the set of methods employed in developing our solution and its results and discussion. The solution we built works in three different levels of hierarchy, tissue type, cancer subtype and genetic mutation classification. One of the main reasons for this hierarchical separation is the distinction in image magnification. Pathologists inspect the tissue samples using various magnifications/zoom levels. Figure 5.1 demonstrates the distinction between 40x and 5x magnification. The former displays the individual characteristics of each cell and the latter displaying more general patterns in a broader overview of the tissue.

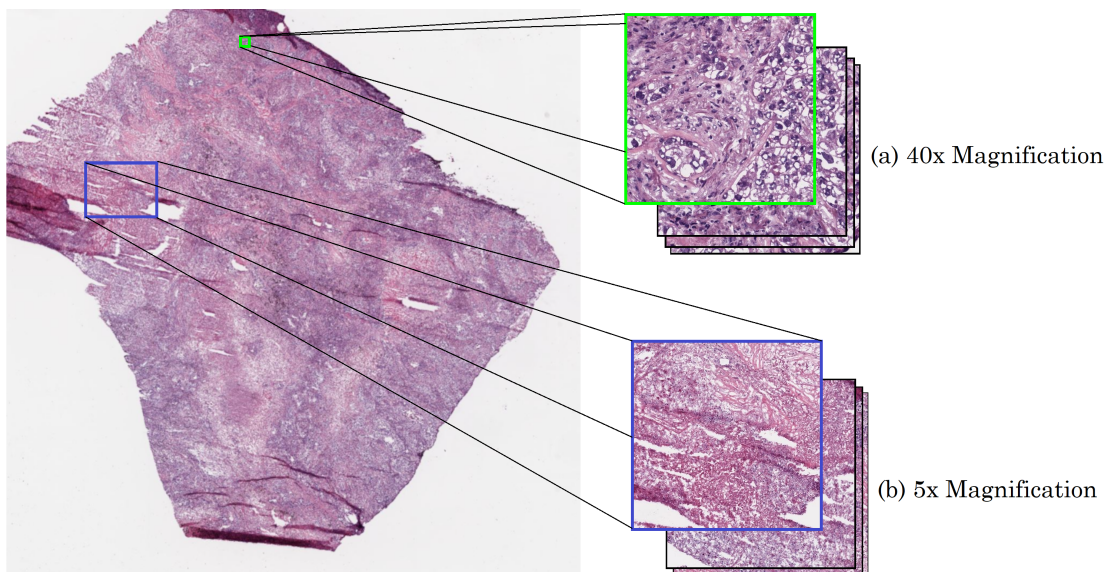


Figure 5.1: Comparison between tiles acquired from 40x and 5x magnifications.

The following sections describe in detail all of the elements composing our pipeline. Furthermore, experiment design and results are presented and discussed for each classification task, identifying the benefits and limitations found.

## 5.1 Tissue Type Classification Model

The first hierarchical level is responsible for the distinction between normal and tumour tissue through the inspection of preprocessed H&E stained histopathological images.

### 5.1.1 Data Gathering and Categorisation

The H&E slides used for training purposes in our classification models have been acquired from the datasets previously identified and described in detail in Chapter 4. This acquisition was made using the Genomics Data Commons (GDC) Data Transfer Tool, a client-based mechanism that allows the download and submission of data from GDC. A set of queries was used to select and download files specific to both TCGA-LUSC and TCGA-LUAD programs, including the associated H&E .svs whole slide images. Each of these images contains a specifically constructed file name following the TCGA-Barcode format. In this format, the file name contains a set of fields specifying the sample's characteristics, including the original project, patient ID and sample type as displayed by Figure 5.2.

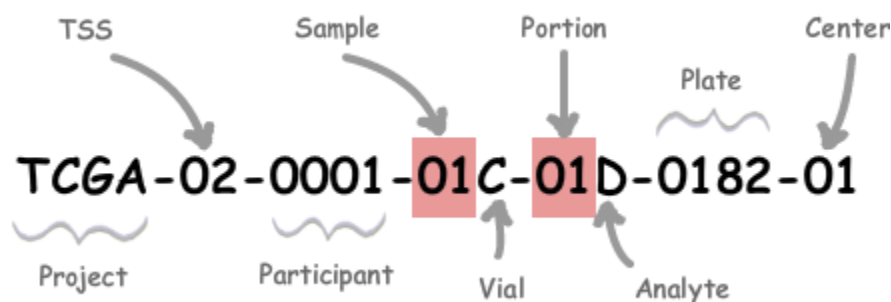


Figure 5.2: TCGA-Barcode file name format example. The "Project" field designates the name of the original project; "TSS" describes the tissues source site, 02 refers to MD Anderson Cancer Center; "Participant" describes the patients' alpha-numeric value, 0001 is the first patient of the study; "Sample" describes the tissue type, 01 for Primary Solid Tumour; "Vial" describes the sample order, C is the third sample from the same patient; "Portion" describes the portion order of the sample, 01 is the first portion; "Analyte" describes the molecular type of analyte for analysis, D is for DNA; "Plate" identifies the order of plate in a sequence of 96-well plates, 0182 is the 182nd plate; "Center" identifies the centre receiving the sample for analysis, 01 represents The Broad Institute GCC.

Once acquired, this file name format aided categorising the images in these datasets. The sample type field in each image allowed us to separate them into cancerous and non-cancerous groups accordingly. The reason for this is that cancerous samples are represented with sample types ranging from 01 to 09, while on the other hand, normal samples have a sample type ranging from 10 to 14. Moreover, files were also split and stored according to their original program and lung cancer subtype, squamous cell and adenocarcinoma, TCGA-LUSC and TCGA-LUAD (see Figure 5.3).

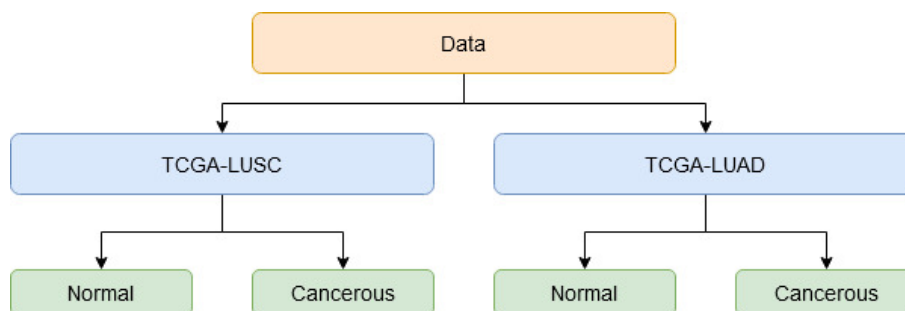


Figure 5.3: Folder structure used to store tissue type classification data.

### 5.1.2 Data Preprocessing

The original H&E whole slide images come in the .svs format and with infeasible dimensions to feed the classification network. Therefore, a set of preprocessing steps were required to turn these images manageable. The first step was to re-format them, and to this end, images were opened with *PIL* framework, transformed into *NumPy* arrays and saved into new PNG files.

Afterwards, the new re-formatted images get divided into several same-sized tiles acquired from a 40x magnification, whose dimensions were specified and set to 192x192. Only a few get selected from these tiles to serve as an input for our classification task. This selection was made with the aid of a scoring function introduced by Erikson *et al.*[15], where each tile gets scored according to their tissue percentage, tissue colour and saturation. The scoring process will attribute each tile a colour based on its utility. Green and red are the edges of the colour spectrum used, where green represents the most informative tiles with score values ranging from 0.8 to 1 by exhibiting high tissue percentages and preferred staining characteristics. In contrast, red is reserved for tiles with a score of 0, containing no tissue at all. Intermediate scores are represented with orange and yellow, with the former ranging from 0 to 0.1 and the latter ranging from 0.1 to 0.8. (see Figure 5.4).



Figure 5.4: Tiling and scoring process; green squares represent the most informative tiles and red squares represent the least informative ones, adapted from Erikson *et al.* [15].

Inline with Erikson *et al.* [15], only the 50 highest scoring tiles of each image get selected and retrieved to represent instances composing the image bag. Every bag previously formed is then categorised into positive and negative classes. The positive class represents cancerous tissue, and the negative one represents normal tissue samples.

### 5.1.3 Baseline Classifier

A set of experiments was conducted to lay the ground for further advances in developing our final model, taking advantage of previously identified superior results present in the state of the art literature review. In this review, the most commonly used models, and the ones that displayed the best accuracy values, were built using CNN architectures such as Google’s Inception V3. Likewise, we built a set of preliminary models employing two different approaches using our previously labelled

and preprocessed images, firstly to fine-tune an Inception V3 model previously trained on the Imagenet dataset [12], and secondly to fully train an Inception V3 model from scratch. The former produced slimmer results, with accuracy and AUC values around 0.9. The latter demonstrated the previously identified potential for CNN's in classification tasks such as the one herein discussed by achieving accuracy and AUC values around 0.96. These values stand reasonably close to the ones produced by the state-of-the-art approaches summarised in our literature review.

Nonetheless, these approaches introduced a somewhat famous theoretical issue in classification tasks based in Machine Learning. The root of this issue is the labelling of the tiles used to train our models. Data samples labelled as belonging to cancerous tissue vary on the extent and percentage of tumour tissue. This means that not every "piece" acquired in the previously described tilling process represents actual cancerous tissue in the original whole slide image.

Considering this issue, the approaches mentioned above turn theoretically flawed, as the input images used may be incorrectly labelled and thus may introduce error in the model. Because this theoretical problem is not recent nor unusual, there is extensive research to formulate new approaches able to deal with it.

#### 5.1.4 Attention-Based Classifier

New experiments were then conducted using one of these approaches, specifically MIL (see Chapter 2.6), using whole slide images as **bags** and the many tiles as **instances** to train a bag-level classifier. When employing this formulation, our model's only concern is to classify the whole image into a positive or negative class, thus never facing the problem identified in previous experiments.

To this extent, a new model was constructed using the MIL approach proposed by Ilse *et al.* [26]. In this particular approach, the model's convolutional layers are followed by a pair of novelty attention mechanisms, proposed to substitute the regular MIL pooling layers commonly used. The rationale behind this substitution is entangled with the "static" characteristics of classical pooling layers. Maximum and mean operations represent the most commonly used pooling layer operations, where the output is generated calculating the maximum or mean values for the inputs. Although efficient, these operations turn the decision making strict, removing the ability to adapt to cases exhibiting unique properties.

Figure 5.5, adapted from Ilse *et al.* [26] demonstrates the distinction between most common pooling layers and the proposed attention mechanism. The first case, Figure 5.5 (a) represents the instance-level approach, and the final MIL pooling layer aggregates the predictions generated for each tile using the typical max or mean operators. The second case, Figure 5.5 (b), represents the bag-level approach and the MIL pooling layer precedes the last fully connected layer. Lastly, Figure 5.5 (c) represents the approach employed in our model, using the proposed attention mechanism before the last fully connected layer to attribute an attention value to each instance.

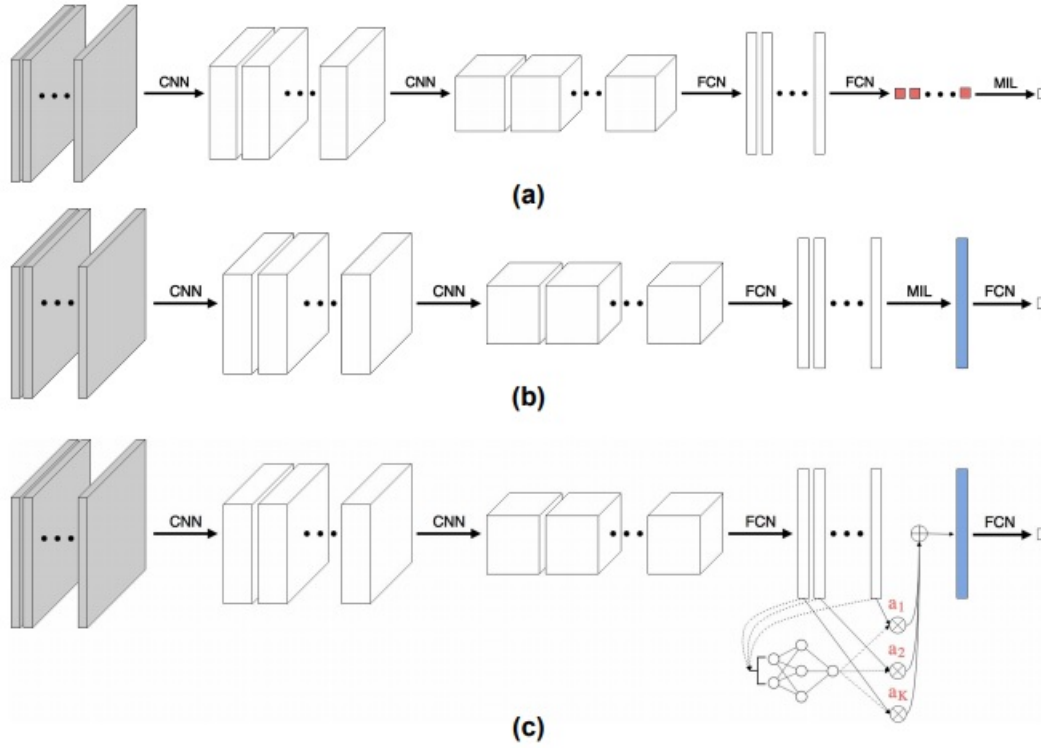


Figure 5.5: Summary of classical pooling techniques (a,b) and the novelty attention mechanism (c), from Ilse *et al.* [26].

The novelty attention mechanisms work by creating a weighted average of instances, where the network determines the weights. Label prediction is formulated in Equation 5.1, where  $z$  is the label attributed to the bag containing  $K$  instances and  $a_k$  is the attention factor attributed to instance  $h_k$ . Equation 5.2 shows the formulation for the Default attention factor, where  $w$  and  $V$  represent two trainable variables. Similarly, Equation 5.3 demonstrates how the Gated attention mechanism is formulated, also containing three trainable weights,  $w$ ,  $V$ , and  $U$ .

$$z = \sum_{k=1}^K a_k h_k \quad (5.1)$$

$$a_k = \frac{\exp\{w^T \tanh(Vh_k^T)\}}{\sum_{j=1}^K \exp\{w^T \tanh(Vh_j^T)\}} \quad (5.2)$$

$$a_k = \frac{\exp\{w^T (\tanh(Vh_k^T) \odot \text{sigm}(Uh_k^T))\}}{\sum_{j=1}^K \exp\{w^T (\tanh(Vh_j^T) \odot \text{sigm}(Uh_j^T))\}} \quad (5.3)$$

The distinction between the two proposed attention mechanisms rises with the introduction of the gating operation,  $\odot \text{sigm}(Uh_j^T)$  in Equation 5.3. According to Ilse *et al.*, the introduction of this operation, initially proposed by Dauphin *et al.* [11], tackles the approximate linearity of the  $\tanh(x)$  function for  $x \in [-1, 1]$  with the sigmoid function non-linearity. This conjugation



creates a non-linear trainable pooling layer capable of capturing small subtleties characterising each class. Additionally, the proposed pooling methods allow the identification of key instances for the whole bag label attribution. Figure 5.6 shows the full pipeline of our solution, aggregating the data collection, preprocessing and categorisation with the training of the Attention-Based model detailed above and classification of new data.

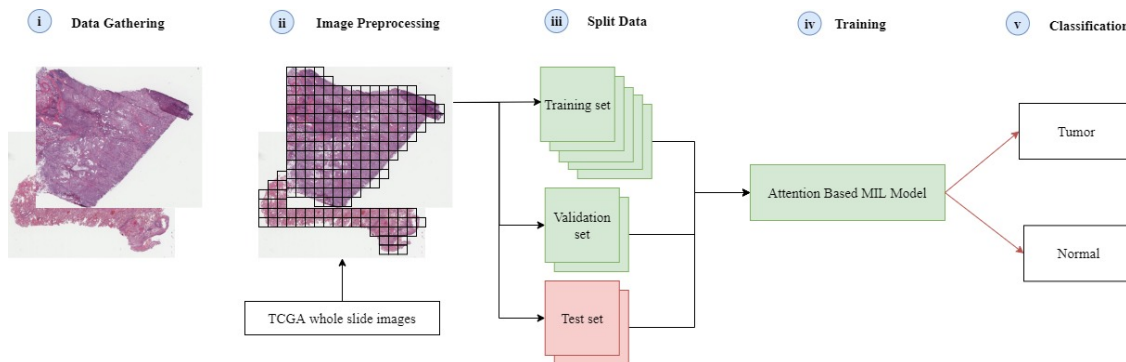


Figure 5.6: Tissue Type Classification Model Pipeline: (i) TCGA data is downloaded (ii) images are tiled and scored according to utility (iii) data is split into training, validation and testing subsets (iv) MIL Attention-Based model, trained with the data (v) classification on test set.

### 5.1.5 Training, Validation and Testing

Our experiment design was delineated to employ a manual grid-search to fine-tune some of the hyper-parameters composing the model. To this end, two hyperparameters were chosen, learning rate and weight decay, and their values transited between  $1e-4$ ,  $5e-4$  and  $1e-3$  in each experiment. All of these experiments employ the Nadam [14] optimizer, using the default values  $\beta_1, \beta_2$  of 0.9 and 0.999.

Furthermore, the K-Fold Cross-Validation technique was used to resample our data. In this resampling technique, the original dataset gets split into same-sized K folds/groups of samples. A new array of K datasets is generated, with each one randomly selecting different folds for training, validation and test phases. This method is particularly interesting as it does not introduce order bias in the model and validates its performance using different test sets each time. Our preprocessed data samples were split into the usual training, validation, test sets using a 3-Fold Cross-Validation method (see Figure 5.6). Thus, we train and evaluate the model three times for each experiment, assuring distinct training and testing samples. A general metric on the model's performance is generated by computing the mean bag-level prediction accuracy (Acc) and Area under the Curve (AUC), using the results from the three runs.

### 5.1.6 Results

The following table summarises the best results acquired in the experimental process described above. Table 5.1 displays the highest Acc and AUC values achieved using the Default Attention

and Gated Attention mechanisms defined in Chapter 5.1.4 (Equations 5.2 and 5.3). Moreover, the hyperparameters used in the top performant experiments are also detailed.

Table 5.1: Best tissue type classification results for each attention mechanism.

Attention Mechanism	LR	Weight Decay	Accuracy	AUC
Standard	1e-4	1e-4	0.900	0.939
Gated	1e-4	5e-4	0.912	0.945

The highest Acc and AUC values achieved belong to the experiments employing the Gated Attention mechanism and using the learning rate and weight decay values of 1e-4 and 5e-4. Moreover, experiments employing the Default Attention mechanism also reached compelling Acc and AUC values, although using a slightly smaller weight decay factor. The slight improvement in results from the Gated Attention mechanism follows the original proposition by Ilse *et al.* [26]

### 5.1.7 Attention Maps

To discern the actual quality of our model's predictions and further aid pathologists in the visual inspection of our whole slide images, we adopted the Grad-Cam [49] algorithm. The Grad-Cam algorithm highlights regions of interest in the original images by inspecting gradient information generated through the network's layers. Moreover, the gradient values tie each neuron to its relative significance in the predictions generated. With this information, a heatmap of "importance" is generated and superimposed over the original image.

An experiment was conducted using a TCGA-LUSC whole slide image labelled as cancerous. This image was then preprocessed and fed as an input to our model, which correctly classified it. Later, the attention maps highlighting the regions contributing to this classification were generated using the GradCam algorithm. Figure 5.7 summarises the experiment's workflow, with the original whole slide image represented on the left side. The scattered green squares represent the selected tiles fed into our fully-trained model. On the right, the Grad-Cam algorithm's superimposed image highlights the regions with a larger contribution to the model's prediction. Recalling the visual features defining cancerous cells presented in Chapter 2.2, one might argue that the algorithm has highlighted cells exhibiting the same anomalies, including variances in nuclei size and shape, reduced cytoplasm and smudged chromatin.

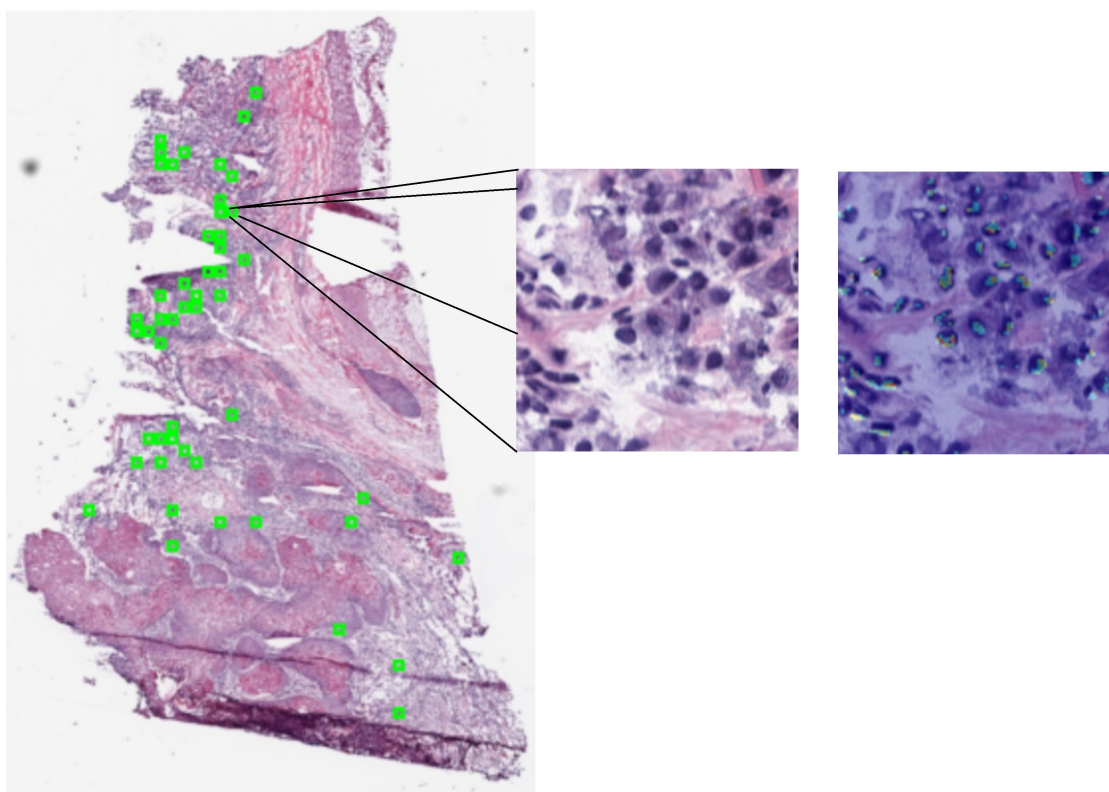


Figure 5.7: Attention Maps Generated with GradCam algorithm, with highlighted atypical cells exhibiting large and abnormally shaped nuclei.

### 5.1.8 Discussion

The developed models exhibit significant potential in the tissue type classification of H&E stained whole slide images, achieving an overall best AUC value of 0.945. Table 5.2 displays the comparison between the results achieved with our models and the results present in the literature.

Table 5.2: Tissue type classification performance comparison.

Model	AUC
Conventinonal Machine Learning, Yu <i>et al.</i> [57]	0.850
SqueezeNet, Li <i>et al.</i> [35]	0.911
HaloAi, Pham <i>et al.</i> [46]	0.922
Our model, default attention mechanism	0.939
Our model, gated attention mechanism	0.945
Inception V3, Coudray <i>et al.</i> [10]	0.993

Our models' ability to distinguish normal and cancer tissue samples was able to surpass several solutions presented in the literature and place it close to the previously described state-of-the-art solution by Coudray *et al.* Although both attention mechanisms achieve compelling performances, a slight improvement is observed in the employment of the Gated Attention mechanism. As previously mentioned, this insight follows the comparative proposition by Ilse *et al.* in their publication.

We argue that this improvement is closely related to the introduction of the *sigmoid()* function non-linearity mentioned in Chapter 5.1.4. This introduction allows the model to increase its ability to identify the small subtleties characterising normal and cancer tissue by reducing the range of values attributed to features belonging to normal tissue. Furthermore, from the results produced in our experimental process, we can gather that the developed models identify the morphological features, mentioned in Chapter 2.2, characterising normal and cancer cells. This idea is further enforced with the results produced by the GradCam algorithm, which highlighted such features in the images that our model predicted as cancerous. These features included large and abnormally shaped nuclei, smudged chromatin and reduced cytoplasm area.

Ultimately, these results enforce the idea that Deep Learning approaches can further aid pathologists in developing a diagnosis. Additionally, a great deal of interpretability is introduced, allowing a more thorough validation and increasing trust in the models. Future work and further research would benefit from the inclusion of new tests using independent cohorts from external sources. Given that the classification is based on a cellular level, experimentation on tissues excised from other organs could also benefit the model's validation, *e.g.* breast, colon or brain cancer tissue. Moreover, on a technical perspective, better results could be achieved with future work using different complexities in model architectures and hyperparameterization techniques.

## 5.2 Lung Cancer Subtype Classification Model

The second level of the hierarchical model is concerned with the identification of the NSCLC subtype present in histopathological images. To this extent, our solution employs the previously formulated pipeline, changing only the input images and corresponding labels. Only the two most common NSCLC subtypes, squamous cell carcinoma and adenocarcinoma, were selected for this new classification task.

### 5.2.1 Data Gathering and Categorisation

The histopathological images used in this second classification task flow from the TCGA-LUSC and TCGA-LUAD datasets previously gathered in the process described in Chapter 5.1. Only cancerous samples are selected in this particular task, as displayed in Figure 5.8. Such selection is employed as normal tissue samples do not carry the visual features characterising each subtype, as previously detailed in Chapter 2.3.

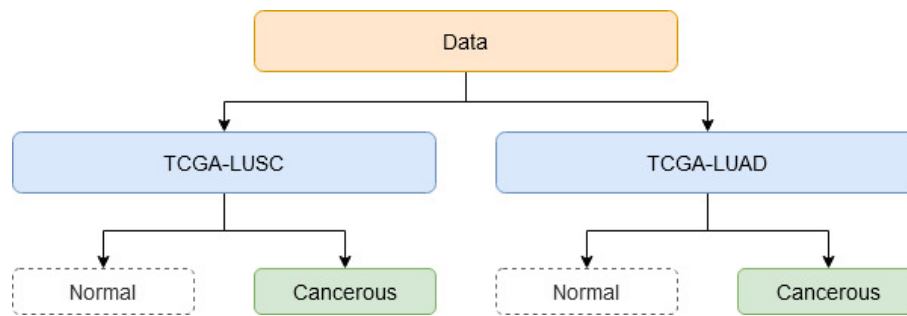


Figure 5.8: Folder structure used to store lung cancer subtype classification data.

### 5.2.2 Data Preprocessing

The preprocessing workflow used in the previous model demonstrated to be less valuable as it specialised in generating tiles from the original images at a very high level of magnification/zoom. The particular patterns characterising NSCLC subtypes cannot be observed at such high magnifications. Thus a new preprocessing workflow was developed using the OpenSlide framework [17] to tile images at lower magnifications. The most commonly used low magnifications are 5x, 10x and 20x, with the first exhibiting the pretended visual features in all of the samples. In contrast, 10x and 20x magnifications often miss these features in images with smaller sizes. Hence, all samples were tiled at a 5x magnification, with each tile measuring 299 pixels in width and height (see Figure 5.9). Similarly to the scoring function proposed by Erikson *et al.* [15], the percentage of tissue in each tile holds considerable relevance. A threshold for tissue percentage was defined at 0.75, and tiles containing less tissue percentage were discarded.

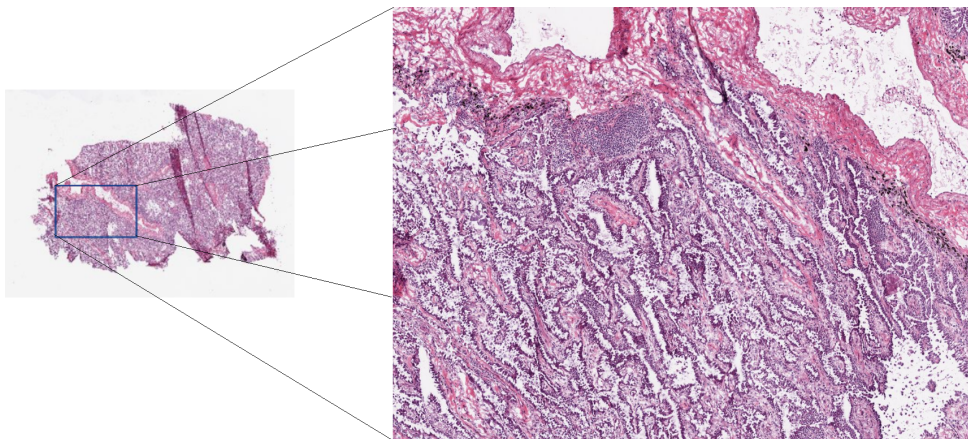


Figure 5.9: TCGA-LUAD sample tile acquired at 5x magnification exhibiting papillary patterns, a visual clue for adenocarcinoma diagnosis.

### 5.2.3 Attention-Based Classifier

No significant changes were applied to our classifiers' pipeline or architecture, defined in section 5.1. However, small tweaks were necessary for the network to deal with the new tile dimensions. These tweaks included the reshaping of input tensors, initially designed to hold 192x192 RGB tuples, into 299x299 RGB tuples. Moreover, the classification model no longer tries to predict the presence of cancer in input images. Instead, it tries to predict the presence of features belonging to adenocarcinoma or squamous cell carcinoma.

### 5.2.4 Training, Validation and Testing

The experiment design employed in this second task was similar to the one employed in tissue type classification, with all experiments following the same structure and employing the Nadam [14] optimisation algorithm. In these experiments, the same manual grid-search approach is followed to analyse the impact of different hyperparameters.

The 3-Fold-Cross-Validation resampling method mentioned in the previous section was also employed to split samples into training, validation and test sets. The results for each experiment gather a general metric, computed using the mean bag-level Acc's and AUC's achieved.

### 5.2.5 Results

The following table, Table 5.3, summarises highest mean Acc and AUC achieved in the many experiments conducted using the two attentions mechanisms studied in our work.

Table 5.3: Best cancer subtype classification results for each attention mechanism.

Attention Mechanism	LR	Weight Decay	Accuracy	AUC
Standard	1e-4	1e-4	0.62	0.780
Gated	1e-4	5e-4	0.65	0.801

The highest classification performance was achieved using the Gated Attention mechanism, with Acc values peaking around 0.65 and AUC value peaking around 0.8. Equivalently to the tissue type classifier presented in the previous chapter, the proposition formulated by Ilse [26] is confirmed with an increase in performance using the Gated Attention mechanism.

### 5.2.6 Attention Maps

In a similar fashion to the experiment described in 5.1.7, the GradCam algorithm was used to validate our models' performance. To do so, we initially fed an image from the TCGA-LUSC dataset to our model, which then correctly predicted its label as squamous cell carcinoma. Afterwards, the GradCam algorithm was used to generate the attention maps highlighting the regions in the image corroborating the prediction.

Figure 5.10 summarises the experiment's workflow, with the original whole slide image represented in the left side. The highlighted blue square in the original whole slide image represents

one of the tiles fed into our model. In this tile, a red oval shape highlights what can arguably be considered a keratin pearl. The formation of these pearls as been demonstrated to be characteristic of squamous cell carcinomas (see Chapter 2.3). On the right side of Figure 5.10, the attention map generated by the Grad-Cam algorithm highlights this pearl as the main contributor to the prediction generated. This highlight enforces the models' ability to discern morphological features characteristic of each subtype of lung cancer.

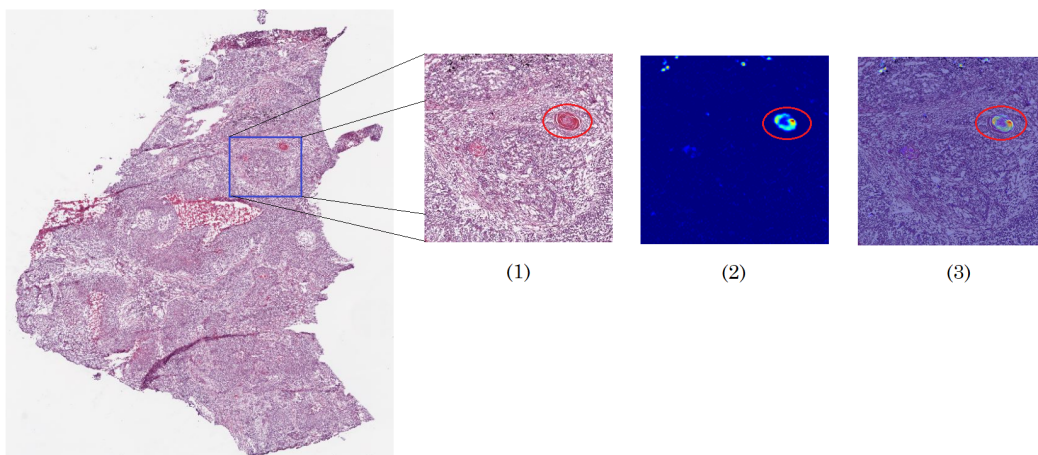


Figure 5.10: Attention Maps Generated with GradCam algorithm, with highlighted keratin pearl formation, (1) original tile; (2) attention map generated; (3) superimposed image.

### 5.2.7 Discussion

Our proposed subtype classification model demonstrated a significant decrease in performance when compared to the previously described tissue type classifier. Nonetheless, it still achieves compelling AUC values peaking around 0.8 and demonstrates its potential by correctly identifying subtype morphological characteristics such as keratin pearls. Table 5.5 promotes a comparative look between the results achieved with our solution and the solutions in the literature.

Table 5.4: Lung cancer subtype classification performance comparison.

Model	AUC
Conventinonal Machine Learning, Yu <i>et al.</i> [57]	0.750
Our model, using default attention mechanism	0.780
Our model, using gated attention mechanism	0.801
Inception V1 Fine-Tuned, Khosravi <i>et al.</i> [29]	0.830
Inception V3, Coudray <i>et al.</i> [10]	0.950

Similarly to our tissue classification model, the results we obtain with our subtype classifiers outperform the conventional Machine Learning solutions. On the other hand, our models are not able to outperform the Deep Learning solutions developed by Khosravi *et al.* and Coudray *et al.*, even though they are closer to the former than the latter.

A set of limitations may justify this performance gap. The first one is directly correlated to the heterogeneity among images belonging to each NSCLC subtype. Aside from the visual features identified in Chapter 2.3, squamous cell carcinoma and adenocarcinoma images remain very similar. This similarity introduces a lot of noise in the network as only a few instances contain the characteristic features with informational gain. Additionally, both solutions developed by Khosravi *et al.* and Coudray *et al.* employ very complex CNN architectures, whereas our models employ a relatively simple one. This distinction in architectures may also justify the models' inability to identify unique morphological features characteristic to each NSCLC subtype. Furthermore, and in agreement with the original paper by Ilse *et al.*, our model's trainable pooling layer may be learning the mean operator, which leads to very similar predictions in both classes, and an inability to distinguish them.

Future work and research should be focused on the employment of novelty architectures, with higher levels of complexity, and multiscale solutions, taking advantage of morphological features at different levels of magnification/zoom [20], that we may not be able to identify when we fix the 5x magnification/zoom for every sample.

### 5.3 Lung Cancer Gene Mutation Classification Model

The third and final level of the hierarchical model is concerned with the identification of genetic mutations in lung cancer patients through the inspection of histopathological images. Equivalently to the previous hierarchical level, this solution employs the previously formulated pipeline, changing only the input images and corresponding labels.

#### 5.3.1 Data Gathering and Categorisation

The patients' genome sequencing data mentioned in Chapter 4.2 was used to label their histopathological images according to their genetic mutations. However, only the mutational status information of each patient was needed for this labelling process. This information was already bundled and published in the GDC through the efforts of the TCGA program. Thus we were able to access it using a set of queries similar to the ones used to acquire the original histopathological images, described in Chapter 5.1.1. This set of queries resulted in a .json file containing each patient ID and the mutational status of each gene. This information was then packed into a dictionary using a Python script, which later labelled each image according to its patients mutational status. Figure 5.11 illustrates the formulated dictionary, with each key representing a patient ID and the associated values representing a set of genes and its mutations status, with 1 representing mutated genes and 0 otherwise.



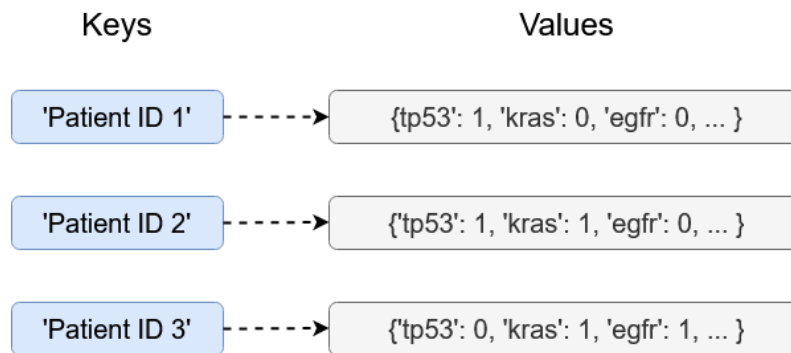


Figure 5.11: Dictionary containing the genetic mutational status for each patient; each key represents a patients ID and the associated values represent the mutational status of each gene, 1 indicates mutated genes, 0 otherwise.

Using the information in this dictionary, the previously gathered histopathological images were then labelled according to their patients mutated genes. To this end, only cancerous samples were selected, as illustrated in Figure 5.12. The final folder structure allows the distinction between histopathological images representing patients with genetic mutations and patients with no genetic alterations.

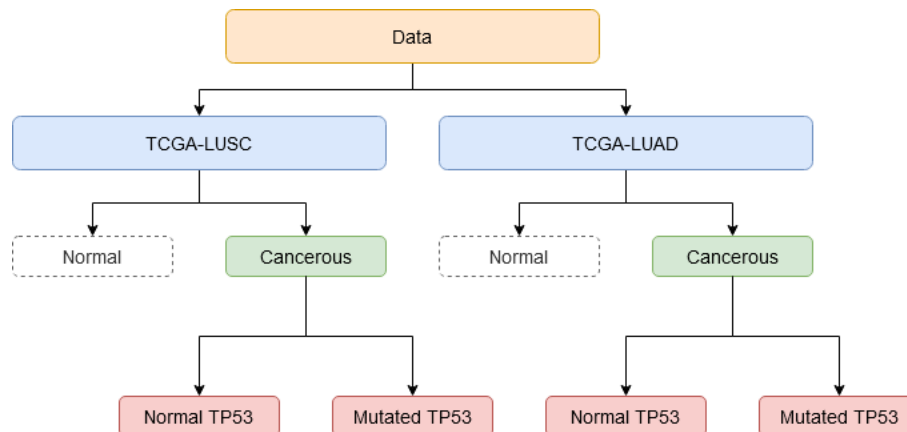


Figure 5.12: Folder structure used to store genetic mutation classification data, e.g. TP53 gene mutation data.

### 5.3.2 Data Preprocessing

Studies and research conducted on the identification of genetic mutations through the inspection of histopathological images have yet to find specific morphological features characterising each mutated gene and thus a correlation between the images magnification and the actual mutation status. However, previous works by Coudray *et al.* [10] and Kim *et al.* [30], described in Chapter 3.2, make use of data samples acquired at 20x magnifications. We argue that this level of magnification and lower ones may have an impact on the observable morphological features we

try to identify. Therefore, we followed a similar path, with our samples being collected at 5x magnification, following the same preprocessing method described in Chapter 5.2.2.

### 5.3.3 Attention-Based Classifier

Equivalently to the previous classification task, no significant changes were necessary to employ in our classifiers' pipeline or architecture. As the tiles acquired from each histopathological image remain in the 299x299 RGB tuple format, no further tweaks were necessary to accommodate the new samples.

Nonetheless, this new classification task is distinct, and it must be noted that the classifiers focus is not the distinction of samples exhibiting morphological features characterising each NSCLC subtype. Instead, the focus is centred on identifying morphological features characterising patients with genetic mutations.

### 5.3.4 Training, Validation and Testing

The experiment design employed in this third task remained similar to the one employed in tissue type and cancer subtype classification, using the same optimisation algorithm and employing the same manual grid-search method to fine-tune hyperparameters.

A single distinction was present in this experiment design, as datasets and experiments were built to accommodate each studied gene. The genetic mutation distributions in Chapter 4.2 were used to select the genes for each dataset and experiment. As research remains slightly slim, we favoured the genes selected by Coudray *et al.* [10], which were identified in the TCGA-LUAD cohort. Thus, using this cohort, we conducted experiments on datasets comprising histopathological images from patients exhibiting TP53 and KRAS mutations and patients with no alterations in these genes.

For each experiment, the 3-Fold-Cross-Validation resampling method was also employed to split samples into training, validation and test sets. The results for each experiment gather a general metric, computed using the mean bag-level Acc's and AUC's achieved.

### 5.3.5 Results

The following tables summarise our models' performances in the identification of genetic mutations. Table 5.5 displays the results regarding TP53 mutations, and Table 5.6 displays the results regarding KRAS mutations.

Table 5.5: Results from TP53 gene mutation prediction in TCGA-LUAD set.

Attention Mechanism	LR	Weight Decay	Accuracy	AUC
Standard	1e-4	1e-4	0.520	0.633
Gated	1e-4	5e-4	0.545	0.650

Table 5.6: Results from KRAS gene mutation prediction in TCGA-LUAD set.

Attention Mechanism	LR	Weight Decay	Accuracy	AUC
Standard	1e-4	1e-4	0.600	0.654
Gated	1e-4	5e-4	0.610	0.671

The highest classification performance was achieved in the prediction of KRAS mutations. This performance was achieved using the Gated Attention mechanism, with Acc values peaking around 0.61 and AUC value peaking around 0.67.

### 5.3.6 Discussion

Our models' manifested a significant decrease in performance when assigned with the task of identifying patients genetic mutations through the inspection of their histopathological images. This decrease in performance was not a surprise, as the previously conducted literature review process pointed this task as very complex and with relatively slim research. In fact, regarding lung cancer, only one major scientific publication has been made describing an approach to this task. This publication was developed by Coudray *et al.*, and in Table 5.7, we compare its results to ours.

Table 5.7: Genetic mutation classification performance comparison.

Model	Gene	AUC
Our model, using gated attention mechanism	TP53	0.650
Our model, using gated attention mechanism	KRAS	0.671
Inception V3, Coudray <i>et al.</i> [10]	TP53	0.760
Inception V3, Coudray <i>et al.</i> [10]	KRAS	0.733

In a similar fashion to the tasks previously described, the solution developed by Coudray *et al.* achieves the state-of-the-art AUC values. However, the gap in results identified in this task is significantly smaller than the one observed in subtype classification.

Furthermore, we argue that these narrow results derive from the lack of in-depth knowledge of the morphological features characterising each genetic mutation. In both our solution and the solution developed by Coudray *et al.*, a set of patterns may have been identified to characterise specific genetic mutations. However, as no visual insights can be recognised, not even by expert pathologists, we cannot infer on improved ways to identify new patterns.

We theorise that our models' complexity may also be a limitation. As mentioned before, the solution developed by Coudray *et al.* uses the Google Inception V3 architecture, whereas our models' architecture is composed of only two convolutional layers.

Future work should be focused on the development of more complex model architectures and hyperparameter optimization methods. Moreover, research would significantly benefit from further experiments using other candidate genes, other cohorts, or other deceases, *i.e.* experiments using the TCGA-LUSC cohort or experiments using histopathological images from breast cancer patients. Additionally, multiscale solutions could be the origin of breakthroughs in this research,

identifying different characteristic patterns at different magnification levels, something that cannot be achieved in solutions like ours where we fix the magnification level tissue inspection.

## **5.4 Summary**

The results achieved by the solutions we have developed demonstrate the excellent potential for Deep Learning in the medical imaging field. Moreover, these results extend the research efforts presented in the literature, leading to some improvements in performance and interpretability.

Furthermore, from the three classification tasks, the first, tissue type classification, as demonstrated to be the one with the most significant potential to reach pathologist-like performances. This potential is closely related to the advances in Deep Learning solutions, generating CNN architectures capable of identifying the characteristic morphological features present in normal and cancer cells. To a certain extent, this potential has also been identified in the classification of specific NSCLC subtypes. However, the heterogeneity observed in histopathological images from each one of these subtypes remains a significant limitation imposed in this task.

On the other end of the spectrum, identifying genetic mutations in histopathological images has been demonstrated to be a problematical task with performance results standing relatively far from precise. Nonetheless, the continuous research and development may eventually lead to better results, generating new knowledge on the morphological features characterising patients enduring genetic mutations, and displaying this knowledge in an interpretable way.

## Chapter 6

# Conclusions and Future Work

Deep Learning solutions' promising strength has been verified to significantly benefit the medical imaging field. Long and extensive research has led to notable advances and development, including in the lung cancer diagnosis using histopathological images.

Our work expands this research by tackling the limitations found in the literature by employing a Multiple Instance Learning approach, which proved to be theoretically relevant in classification tasks like the ones discussed in our work. Additionally, our work introduces interpretability in our model's predictions, allowing pathologists to validate our models' performance and aiding them in diagnostics, redirecting them to specific regions of interest in the original images.

The experiments conducted throughout our work also allowed us to identify our pipelines' limitations. The somewhat unfortunate results in subtype and genetic classification hold an excellent discussion and research opportunity. For example, modern multi-scale MIL solutions hold a great deal of potential in the field by introducing the ability to retain information from different magnification/zoom levels, turning the model's training and validation processes similar to that of a pathologist.



# References

- [1] Yoshinori Abe and Nobuyuki Tanaka. The hedgehog signaling networks in lung cancer: The mechanisms and roles in tumor progression and implications for cancer therapy. *BioMed Research International*, 2016, 12 2016.
- [2] Sophia Apple. Sentinel lymph node in breast cancer: Review article from a pathologist’s point of view. *Journal of pathology and translational medicine*, 50, 01 2016.
- [3] Matthew H Bailey, Collin Tokheim, Eduard Porta-Pardo, Sohini Sengupta, Denis Bertrand, Amila Weerasinghe, Antonio Colaprico, Michael C Wendl, Jaegil Kim, Brendan Rendon, et al. Comprehensive characterization of cancer driver genes and mutations. *Cell*, 173(2):371–385, 2018.
- [4] Andrew H. Beck, Ankur R. Sangoi, Samuel Leung, Torsten O. Marinelli, Robert J. and-Nielsen, Marc J. van de Vijver, Robert B. West, and Daphne van de Rijn, Matt andKoller. Systematic analysis of breast cancer morphology uncovers stromal features associated with-survival. *Science Translational Medicine*, 3(108):108ra113–108ra113, 2011.
- [5] Dmitrii Bychkov, Nina Linder, Riku Turkki, Stig Nordling, Panu E Kovanen, Clare Verrill, Margarita Walliander, Mikael Lundin, Caj Haglund, and Johan Lundin. Deep learning based tissue analysis predicts outcome in colorectal cancer. *Scientific reports*, 8(1):1–11, 2018.
- [6] Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329 – 353, 2018.
- [7] Anne E Carpenter, Thouis R Jones, Michael R Lamprecht, Colin Clarke, In Han Kang, Ola Friman, David A Guertin, Joo Han Chang, Robert A Lindquist, Jason Moffat, et al. Cell-profiler: image analysis software for identifying and quantifying cell phenotypes. *Genome biology*, 7(10):R100, 2006.
- [8] Yong Chen, Jian-Xin Shi, Xu-Feng Pan, Jian Feng, and Heng Zhao. Identification of candidate genes for lung cancer somatic mutation test kits. *Genetics and Molecular Biology*, 36:455 – 464, 00 2013.
- [9] François Chollet. *Deep Learning with Python*. Manning, 2017.
- [10] Nicolas Coudray, Paolo Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine*, 24, 10 2018.
- [11] Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks, 2017.

- [12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.
- [13] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1):31 – 71, 1997.
- [14] Timothy Dozat. Incorporating nesterov momentum into adam. 2016.
- [15] Deron Eriksson and Fei Hu. Whole-slide image preprocessing in Python, 2018.
- [16] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.
- [17] Adam Goode, Benjamin Gilbert, Jan Harkes, Drazen Jukic, and Mahadev Satyanarayanan. Openslide: A vendor-neutral software foundation for digital pathology. *Journal of pathology informatics*, 4:27, 09 2013.
- [18] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- [19] H. Greenspan, B. van Ginneken, and R. M. Summers. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, 35(5):1153–1159, 2016.
- [20] Noriaki Hashimoto, Daisuke Fukushima, Ryoichi Koga, Yusuke Takagi, Kaho Ko, Kei Kohno, Masato Nakaguro, Shigeo Nakamura, Hidekata Hontani, and Ichiro Takeuchi. Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [21] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.
- [22] Y. He, J. Wei, S. Che, S. Liu, and P. Luo. Computer-aided pathological annotation framework: A deep learning-based diagnostic algorithm of lung cancer. In *2019 International Conference on Information Technology and Computer Application (ITCA)*, pages 110–113, Dec 2019.
- [23] Achim Hekler, Jochen Utikal, Alexander Enk, Wiebke Solass, Max Schmitt, Joachim Klode, Wiebke Sondermann, Cindy Franklin, Felix Bestvater, Michael Flaig, Dieter Krahl, Christof Kalle, Stefan Fröhling, and Titus Brinker. Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. *European Journal of Cancer*, 118:91–96, 07 2019.
- [24] Le Hou, Dimitris Samaras, Tahsin M Kurc, Yi Gao, James E Davis, and Joel H Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2424–2433, 2016.
- [25] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and < 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.



- [26] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. 02 2018.
- [27] NIH National Cancer Institute. Cancer Facts & Figures 2020. *CA: A Cancer Journal for Clinicians*, pages 1–76, 2020.
- [28] Km Islam, Trisari Anggondowati, P. Deviany, J. Ryan, A. Fetrick, D. Bagenda, M. Copur, A. Tolentino, I. Vaziri, H. McKean, S. Dunder, J. Gray, C. Huang, and Apar Ganti. Patient preferences of chemotherapy treatment options and tolerance of chemotherapy side effects in advanced stage lung cancer. *BMC Cancer*, 19, 12 2019.
- [29] Pegah Khosravi, Ehsan Kazemi, Marcin Imielinski, Olivier Elemento, and Iman Hajira-souliha. Deep convolutional neural networks enable discrimination of heterogeneous digital pathology images. *EBioMedicine*, 27:317–328, 2018.
- [30] Randie H Kim, Sofia Nomikou, Zarmeena Dawood, George Jour, Douglas Donnelly, Una Moran, Jeffrey S Weber, Narges Razavian, Matija Snuderl, Richard Shapiro, et al. A deep learning approach for rapid mutational screening in melanoma. *bioRxiv*, page 610311, 2019.
- [31] Harrison Kinsley and Daniel Kukiela. *Neural Networks from Scratch in Python*. 2020.
- [32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems*, 25, 01 2012.
- [33] Wayne W. LaMorte. Characteristics of Cancer Cells, 2016.
- [34] Yann Lecun, Leon Bottou, Y. Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278 – 2324, 12 1998.
- [35] Zhang Li, Zheyu Hu, Jiaolong Xu, Tao Tan, Hui Chen, Zhi Duan, Ping Liu, Jun Tang, Guoping Cai, Quchang Ouyang, Yuling Tang, Geert Litjens, and Qiang Li. Computer-aided diagnosis of lung carcinoma using deep learning - a pilot study. 03 2018.
- [36] Haotian Liao, Yuxi Long, Ruijiang Han, Wei Wang, Lin Xu, Mingheng Liao, Zhen Zhang, Zhenru Wu, Xuequn Shang, Xuefeng Li, Jiajie Peng, Kefei Yuan, and Yong Zeng. Deep learning-based classification and mutation prediction from histopathological images of hepatocellular carcinoma. *Clinical and Translational Medicine*, 06 2020.
- [37] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *CoRR*, abs/1312.4400, 2014.
- [38] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafourian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.
- [39] Yun Liu, Krishna Gadepalli, Mohammad Norouzi, George E Dahl, Timo Kohlberger, Aleksey Boyko, Subhashini Venugopalan, Aleksei Timofeev, Philip Q Nelson, Greg S Corrado, et al. Detecting cancer metastases on gigapixel pathology images. *arXiv preprint arXiv:1703.02442*, 2017.
- [40] S. . B. Lo, S. . A. Lou, Jyh-Shyan Lin, M. T. Freedman, M. V. Chien, and S. K. Mun. Artificial convolution neural network techniques and applications for lung nodule detection. *IEEE Transactions on Medical Imaging*, 14(4):711–718, Dec 1995.

- [41] Nicholas J. Mackintosh. Pavlov and associationism. *The Spanish Journal of Psychology*, 6(2):177–184, 2003.
- [42] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.
- [43] Marvin Minsky and Seymour Papert. *Perceptrons; an introduction to computational geometry*. MIT Press, Cambridge, Mass, 1969.
- [44] Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519, 2012.
- [45] Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543–550, 2014.
- [46] Hoa Hoang Ngoc Pham, Mitsuru Futakuchi, Andrey Bychkov, Tomoi Furukawa, Kishio Kuroda, and Junya Fukuoka. Detection of lung cancer lymph node metastases from whole-slide histopathologic images using a two-step deep learning approach. *The American journal of pathology*, 189(12):2428–2439, 2019.
- [47] Talha Qaiser, Yee-Wah Tsang, David Epstein, and Nasir Rajpoot. Tumor segmentation in whole slide images using persistent homology and deep convolutional features. In María Valdés Hernández and Víctor González-Castro, editors, *Medical Image Understanding and Analysis*, pages 320–329, Cham, 2017. Springer International Publishing.
- [48] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.
- [49] Ramprasaath Rs, Michael Cogswell, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. pages 618–626, 10 2017.
- [50] O. Sertel, J. Kong, H. Shimada, U.V. Catalyurek, J.H. Saltz, and M.N. Gurcan. Computer-aided prognosis of neuroblastoma on whole-slide images: Classification of stromal development. *Pattern Recognition*, 42(6):1093 – 1103, 2009. Digital Image Processing and Pattern Recognition Techniques for the Detection of Cancer.
- [51] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*, 09 2014.
- [52] Andreas Stang, Hermann Pohlabein, Klaus M. Müller, Ingeborg Jahn, Klaus Giersiepen, and Karl-Heinz Jöckel. Diagnostic agreement in the histopathological evaluation of lung cancer tissue in a population-based case-control study. *Lung Cancer*, 52(1):29 – 36, 2006.
- [53] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, June 2015.
- [54] William D Travis, Elizabeth Brambilla, H Konrad Muller-Hermelink, and Curtis C Harris. World health organization classification of tumours. *Pathology and genetics of tumours of the lung, pleura, thymus and heart*, 10:179–84, 2004.
- [55] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*, 2016.

- [56] Juanying Xie, Ran Liu, Joseph Luttrell, and Chaoyang Zhang. Deep learning based analysis of histopathological images of breast cancer. *Frontiers in Genetics*, 10:80, 2019.
- [57] Kun-Hsing Yu, Ce Zhang, Gerald Berry, Russ Altman, Christopher Ré, Daniel Rubin, and Michael Snyder. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature Communications*, 7:12474, 08 2016.
- [58] Min Zheng. Classification and pathology of lung cancer. *Surgical Oncology Clinics*, 25(3):447–468, 2016.