Politecnico di Torino

Politecnico di Torino

Computer Engineering Master's Degree

A.a. 2025/2026

Graduation Session March 2026

Evaluating Supervised and Weakly Supervised Learning for the Classification of Sarcomatoid Mesothelioma and Pleuritis

Supervisors:

Francesco Ponzio

Santa Di Cataldo

Candidate:

Samuele Tallone

Abstract

The distinction between pleuritis and sarcomatoid mesothelioma represents an extremely complex challenge in pathological anatomy due to the morphological similarity between the two conditions.

Pleuritis is an inflammation of the pleura, often caused by infections, trauma, or other medical conditions, whereas sarcomatoid mesothelioma is a rare and aggressive form of cancer affecting the pleura, often associated with asbestos exposure. Correct diagnosis is crucial to ensure adequate treatment, as the two conditions have very different prognoses and therapeutic approaches. Histopathological analysis is currently based on Whole Slide Images (WSIs) and requires careful evaluation by expert pathologists, who must thoroughly examine cellular and tissue characteristics to distinguish between these two entities. However, given the extreme histomorphological similarity, this assessment can be subjective and lead to diagnostic errors.

In this context, Artificial Intelligence (AI) emerges as a promising solution to improve diagnostic accuracy, support pathologists in their evaluation, and provide objective metrics to distinguish between different pathologies. Most works in the literature focus on more common diseases, which have larger datasets and often have precise annotations at tissue or cell level made by expert pathologists, enabling the training of supervised models. Conversely, the distinction between pleuritis and sarcomatoid mesothelioma lacks public datasets of this type and has received comparatively less research attention.

In this thesis, we propose various approaches based on AI, using state-of-the-art Deep Learning (DL) techniques, to address the challenge of distinguishing between pleuritis and sarcomatoid mesothelioma. We compare two learning paradigms: a supervised one, leveraging tissue-level annotations, and a weakly supervised one, based on slide-level annotations, which are more readily available. For the weakly supervised approach, based on Multiple Instance Learning (MIL) techniques, we evaluate different combinations of feature extractors and aggregators to identify the most effective configuration for this specific diagnostic challenge. The dataset used contains samples collected from two hospitals in Turin, Molinette and San Luigi, and the slides are stained with Hematoxylin and Eosin (H&E). The dataset closely reflects the real-world scenario of this diagnostic challenge, since it contains a limited number of samples, and only a small number of them is annotated at the tissue level, while the majority of them is annotated at the slide-level. The results obtained demonstrate the potential of AI-based approaches, especially MIL-based ones, to improve the accuracy of diagnosis in this challenging context, since evaluated state-of-the-art feature extractors are capable of extracting discriminating features,

and the MIL-based aggregators are able to effectively leverage them. The ability of the feature aggregators to provide heatmaps, thanks to attention mechanisms, also allows to provide interpretability to the models, which is crucial to support pathologists in their evaluation.

Acknowledgements

Arrivando alla fine di questo percorso, sono soddisfatto di quanto ho realizzato. Sono fiero di me stesso per essere arrivato fino a qui e per non aver mai mollato, anche quando le pressioni erano molte e mi sembrava di non esserne all'altezza.

Mentirei, però, se dicessi che questo traguardo è solo merito mio e che l'ho raggiunto con le mie sole forze. Nei momenti difficili, c'è sempre stato qualcuno al mio fianco a darmi la carica necessaria.

La tesi è stata sicuramente la parte più complessa, e sono grato alla Professoressa Santa Di Cataldo e al Professor Ponzio per avermi guidato in questo percorso. Con la vostra competenza e dedizione siete stati per me un esempio da seguire, e sono onorato di aver potuto apprendere da voi.

Ringrazio anche il Dott. Seyed Mohammad Mehdi Hosseini. Sei stato sempre incredibilmente gentile e paziente; il tuo supporto è stato indispensabile per portare a termine questo lavoro. Nei momenti in cui le cose non andavano come sperato, sei sempre riuscito a rassicurarmi e a infondermi la tranquillità di cui avevo bisogno. Sei una persona estremamente in gamba e di una bontà rara. Ti auguro tutto il meglio che la vita può offrire, sia sul piano professionale sia su quello personale.

Un ringraziamento speciale va ai miei genitori, perché è grazie a voi se ho potuto realizzare il mio sogno. Siete sempre stati al mio fianco, avete supportato ogni mia decisione e mi avete aiutato con tutti i mezzi a vostra disposizione. Non ho mai dato nulla per scontato: so quanto avete sacrificato per me e vi prometto che non sprecherò il futuro che mi avete aiutato a costruire. Siete sempre stati i miei fan numero uno, dai giorni in cui venivate a guardarmi giocare a basket a quelli in cui correvate di fianco a me sugli sci per spronarmi a non mollare. Non riuscirò mai a mettere a parole quanto vi voglio bene e quanto vi sono grato per tutto ciò che avete fatto per me. Spero che, in questo momento, possiate essere orgogliosi di me.

Grazie anche a mia sorella, sempre gentile e paziente, anche quando tornavo da una sessione andata male ed ero scorbutico. Grazie per avermi ascoltato e per essermi stata vicina quando avevo bisogno di sfogarmi.

Ringrazio la mia ragazza Doruntina, che è stata al mio fianco durante l'intero percorso di studi. Chissà quante volte hai dovuto ascoltarmi dire che non ce l'avrei mai fatta: ogni sessione sembrava un ostacolo insormontabile. Alla fine, come

sempre, avevi ragione tu. Gli ultimi mesi non sono stati facili; sei stata spesso la spalla su cui piangevo e, nonostante ciò, non mi hai mai deriso per le mie fragilità. Sono felice che abbiamo completato questo percorso e non vedo l'ora di affrontare le sfide che il futuro ci riserverà. Sono convinto che insieme possiamo fare grandi cose.

Ringrazio i miei nonni e i miei parenti, che mi hanno accudito quando i miei genitori non potevano e che mi hanno visto crescere, aiutandomi lungo il cammino.

Ringrazio infine i miei amici, sia quelli che ho incontrato ogni giorno in università sia quelli che mi hanno supportato da lontano. Siete stati fondamentali per me. Mi avete incoraggiato a proseguire, e la vostra presenza mi ha sempre motivato a migliorarmi. Quando le cose andavano bene, gioivamo insieme; quando andavano male, le affrontavamo sempre insieme. Ovunque ci porterà la vita, sono certo che troveremo sempre il modo di ritrovarci per una delle nostre conversazioni serali, quelle che ti fanno andare a dormire con un sorriso. Vi voglio bene: siete persone straordinarie e sono sicuro che il futuro vi riserverà grandi soddisfazioni.

Dal profondo del mio cuore, grazie a tutti.

Table of Contents

List of Tables

List of Figures

Acronyms

Convolutional Neural Network (CNN)

Artificial Intelligence (AI)

Machine Learning (ML)

Deep Learning (DL)

Support Vector Machine (SVM)

Whole Slide Image (WSI)

Multiple Instance Learning (MIL)

Hematoxylin and Eosin (H&E)

Self-Supervised Learning (SSL)

Weakly-Supervised Learning (WSL)

Deep Neural Network (DNN)

Vision Transformer (ViT)

Region of Interest (ROI)

Microns Per Pixel (MPP)

Fluorescence In Situ Hybridization (FISH)

Pyramid Position Encoding Generator (PPEG)

Transformer-based Patch Transformer (TPT)

Multi-head Self-Attention (MSA)

Enhanced Positional Encoding Generator (EPEG)

Masked Image Modeling (MIM)

Automated Slide Analysis Platform (ASAP)

Chapter 1

Introduction

1.1 Digital Pathology

Digital pathology has revolutionized the field of histopathology by enabling the digitization, storage, and analysis of high-resolution images of tissue samples [1]. This new paradigm moves on from the traditional method of examining glass slides under a microscope, allowing pathologists to view and analyze digital images on a computer screen. These images are known as Whole Slide Images (WSIs) and they are typically acquired using a specialized scanner that captures the entire tissue section at multiple magnifications, allowing for the examination of both low-level structures and high-level details. Having access to high-resolution digital images opens up new possibilities for computer-aided diagnosis, telepathology, and research, since data can be shared more easily among different institutions, studies are more reproducible and computational techniques can be applied to extract quantitative information from the images. The differences between traditional pathology workflow and digital pathology workflow are illustrated in Figure 1.1.

The availability of digital slides has driven a growing interest in applying Artificial Intelligence (AI) techniques to digital pathology, with the goal of improving diagnostic accuracy and reducing the workload for pathologists. There are many studies in this field but, at the time of writing, most of them focus on common pathologies such as breast cancer and prostate cancer, as these have a high incidence rate and large annotated datasets are available for training AI models. To the best of our knowledge, only one study addresses the differentiation between sarcomatoid mesothelioma and pleuritis [2], which is the topic of this thesis. This scarcity of work is due to the difficulty of obtaining the necessary data, since both pathologies have a low incidence rate and the annotation of histopathological images requires significant time and expert knowledge.

This thesis therefore explores two learning paradigms: supervised and weakly

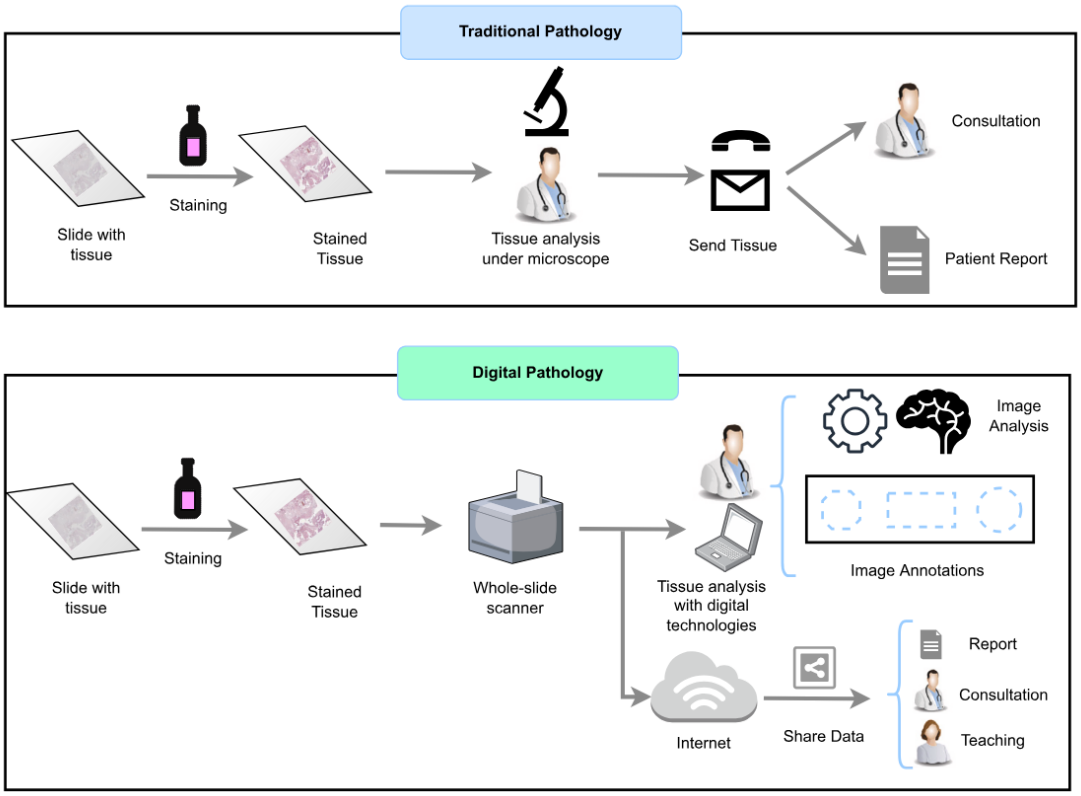

Figure 1.1: Differences between traditional pathology workflow and digital pathology workflow. It is possible to see how digital pathology enables new functionalities such as computer-aided diagnosis, and the possibility for remote, and multi-expert consultations. Whereas in traditional pathology the glass slides must be physically transported to different locations for consultations or second opinions [1].

supervised, by training and evaluating state-of-the-art models on a dataset that is limited in size and only partially annotated, conditions that closely reflect what is generally available in a real hospital setting. The goal is to assess which paradigm is better suited to this challenging scenario, involving two pathologies that are difficult to differentiate even for experienced pathologists.

1.2 Goal

The objective of this thesis is to evaluate the application of artificial intelligence in the field of pathology, specifically by comparing the performance of different state-of-the-art AI models and paradigms on the task of distinguishing sarcomatoid mesothelioma from pleuritis using Hematoxylin and Eosin (H&E) stained WSIs. The comparison is carried out on a dataset with limited size, only partial annotations, and class imbalance, reflecting a realistic hospital scenario. This is a challenging objective from the histopathological point of view, since both conditions share similar morphological features, with pleural cells exhibiting a comparable spindle-shaped appearance. This similarity makes the distinction challenging even for experienced pathologists. A correct diagnosis is, however, extremely important because the two diseases have very different prognoses. Sarcomatoid mesothelioma accounts for 10–20% of all mesothelioma diagnoses and is the most aggressive subtype, capable of metastasizing rapidly to distant organs, with an average prognosis of approximately six months. Pleuritis, on the other hand, is a pleural inflammation that can be caused by a variety of factors but is generally benign.

Another key motivation for this work is the scarcity of available literature on this specific classification task. Other histopathological classification tasks, such as distinguishing mesothelioma subtypes, have received more attention, with multiple papers exploring a variety of AI models and paradigms. In contrast, to the best of our knowledge, only one study has addressed the differentiation between sarcomatoid mesothelioma and pleuritis. The paper was published in 2021 by J. R. Naso et al. [2] and presents a supervised approach that leveraged a large set of patch-level annotated images. While their results are excellent, such annotations are rarely available in practice, as the annotation process must be carried out by experienced pathologists and is extremely time-consuming. The most common scenario, and the one encountered in this work, is one where only a small portion of the dataset contains annotated tissue regions, while the remainder is labeled only at the slide level. Furthermore, the dataset used in this thesis is limited in size, as sharing sensitive patient data for research purposes is often restricted, and is also heavily imbalanced, with differing numbers of samples per class and different acquisition characteristics between the two classes, including magnification level, acquisition year, and hospital of origin.

This thesis therefore aims to expand the available literature on this topic by evaluating different learning paradigms on a dataset that more closely reflects real-world clinical conditions. In particular, this work investigates whether weakly supervised approaches can provide advantages over fully supervised methods when annotations are scarce.

The central research question addressed in this thesis is the following:

> *Do weakly supervised learning approaches outperform fully supervised methods in the task of differentiating sarcomatoid mesothelioma from pleuritis using H&E stained WSIs, when only limited and partially annotated data are available?*

By addressing this research question, we hypothesize that Multiple Instance Learning (MIL)–based approaches can achieve better performance than fully supervised methods when only limited annotated data are available. In particular, MIL frameworks are expected to better exploit slide-level labels by learning discriminative patterns directly from collections of patches without requiring extensive region-level annotations. Furthermore, we hypothesize that foundation models pretrained on large histopathology datasets can generate richer and more generalizable feature representations, which may improve classification performance for this challenging diagnostic task.

1.3 Thesis structure

This thesis is structured as follows:

Chapter 2 provides the theoretical background necessary to understand this work, covering the structure and processing of WSIs, the pathological overview of pleuritis and sarcomatoid mesothelioma, the artificial intelligence paradigms employed, and the software tools and libraries used;

Chapter 3 describes the dataset used in this work, including its composition, the external cohort and the annotations available;

Chapter 4 presents the supervised methodology, detailing the data preprocessing pipeline, the patch-level classifier architecture, and the slide-level aggregation strategy;

Chapter 5 describes the weakly supervised methodology, covering the data preprocessing pipeline, the feature extractors evaluated, and the MIL aggregation architectures employed;

Chapter 6 presents the experimental results, comparing the performance of the different models and paradigms;

Chapter 7 concludes the thesis by summarizing the main findings, discussing their implications, and suggesting potential directions for future research.

Chapter 2

Background knowledge

2.1 Whole Slide Images

WSIs form the basis of digital pathology, as they enable the computational analysis of tissue samples. They represent the digitalized form of the glass slides that are used in traditional pathology [3]. They are created by taking a stained tissue sample on glass and using slide scanners, which utilize advanced optics and digital cameras to scan the entire slide surface. These scanners capture images of each field of view and stitch them together to form a complete, seamless digital representation of the slide. The resulting digital slide can then be viewed, manipulated, and analyzed using dedicated software. This workflow allows for some advantages compared to traditional pathology. First of all, digital slides allow for a high degree of repeatability since they don't require a microscope inspection of the glass slide every time an analysis is necessary, this is particularly useful in large-scale studies. They also facilitate the transfer of specimens for diagnostic purposes, making it easier to have multiple opinions on the same sample. Storage of digital slides is also easier and does not suffer from degradation as traditional glass slides do, even though it comes with its own challenges, given the data sensitivity and the large file sizes of WSIs.

However, the acquisition process of WSIs for AI use is not simple and involves several critical aspects that can be divided into three main categories: pre-acquisition, acquisition, and post-acquisition [3].

Pre-acquisition

In order to have enough images for training AI models it is necessary to include a variety of cases, both neoplastic and non-neoplastic, but many samples are limited in size and quality, reducing the quantity of usable information for model training [3, 4]. In some cases, the tissue samples may be large enough but the relevant area

of interest small, making annotations time consuming and complex. Many times the samples also contain non pathological tissues, or artifacts that can confuse the model during training [3, 5].

Acquisition

During scanning, artifacts present on the glass slide directly affect the quality of the WSI, negatively impacting both the diagnosis and the AI applications [6, 7]. Some common artifacts include:

- **Tissue folding:** when the tissue is not properly flattened on the slide, it can create folds that obscure important details.
- **Air bubbles:** trapped air bubbles can create bright spots that interfere with image analysis.
- **Staining inconsistencies:** variations in staining can lead to color differences that affect the model's ability to generalize.
- **Focus issues:** if the scanner does not maintain proper focus across the entire slide, some areas may appear blurry.
- **Incomplete or improper scanning:** if the scanner misses parts of the slide or does not stitch images correctly, it can lead to gaps or misalignment in the WSI.
- **Dust and debris:** particles on the slide can create artifacts that confuse image analysis algorithms.

The visual examples of some of these artifacts are shown in Figure 2.1.

The characteristics of the scanner itself also play a crucial role in determining the quality of the WSI. Factors such as the sensor used and the light source can affect the image quality, the resolution and the presence of artifacts. Different scanners may also produce images with varying color profiles, which can impact the performance of AI models trained on data from a specific scanner.

Post-acquisition

Once digitized, WSIs require substantial storage capacity and infrastructure, leading to significant economic and organizational costs. The large file sizes require substantial storage capacity, long-term data protection systems, and adequate IT infrastructure.

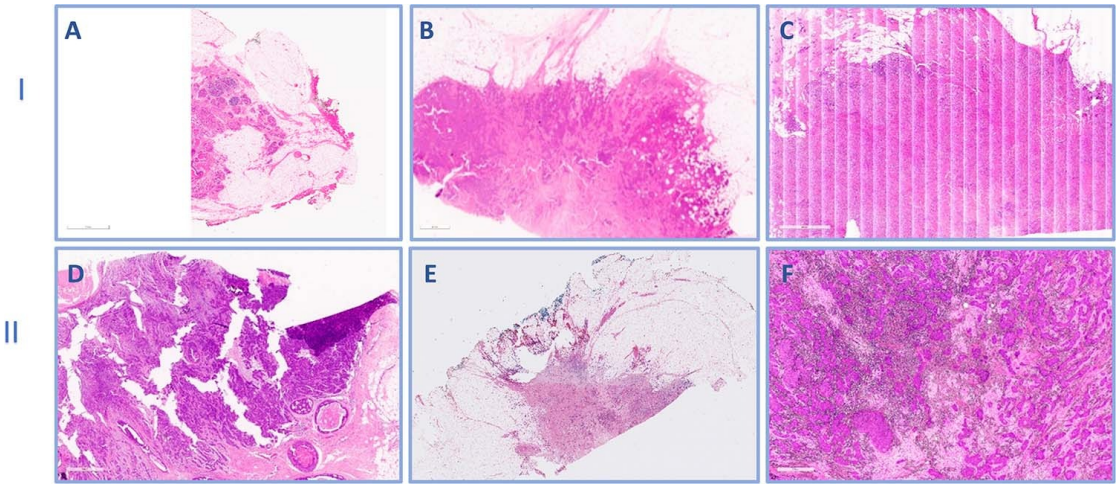

Figure 2.1: Examples of failed WSIs considering quality control. A) Incomplete slide scanning, B) Out-of-focus image, C) Improper line stitching, D) Thick sections with tissue cracking and folding, E) Uneven H&E stain distribution, F) Air bubbles on slide [8, 9].

The key advantage of WSIs, in this context, is that they allow for the use of Machine Learning (ML) algorithms and other computational tools, making it possible to automate the diagnosis process or at least provide insights to the pathologist that can help them in their evaluation. Their use can still provide challenges though, since the size of the images makes it difficult to process them directly with traditional ML algorithms, which usually expect smaller images as input. For these reasons, specific techniques and architectures have been developed to handle WSIs effectively.

2.1.1 Staining methods

As previously stated glass slides are stained before being scanned, this is because cells are usually clear and see-through, making it difficult to study tissue samples under a microscope without staining them. Typically, staining involves using a main dye that colors certain parts of the cell brightly, along with another dye that colors the rest a different color. By using stains, the different parts of the cells react chemically, which allows doctors to better see the tissue's shape, spot any problems, and determine what is happening with a disease.

In the next section is described the staining method that was used for the slides of our dataset:

Hematoxylin and Eosin (H&E) staining

H&E staining represents one of the most widely used staining techniques in histology and is considered the gold standard [10]. This coloration method combines two complementary dyes: hematoxylin and eosin, which together allow a clear visualization of tissue morphology and cellular architecture.

Hematoxylin is a basic dye that binds to acidic structures within the cell, staining them a blue to purple color. It primarily highlights cell nuclei due to its strong affinity for nucleic acids such as DNA and RNA, making nuclear features such as size, shape, and chromatin organization, easily distinguishable under the microscope [10].

Eosin, in contrast, is an acidic dye that binds to basic components of the tissue, staining them in varying shades of pink to red. It mainly stains the cytoplasm and elements of the extracellular matrix, including collagen and connective tissue fibers, providing a counterstain to the hematoxylin-stained nuclei [10]. The combined use of these two dyes generates contrast that makes it easier to distinguish different tissue components and facilitates both qualitative assessment by pathologists and quantitative analysis in digital pathology and AI applications.

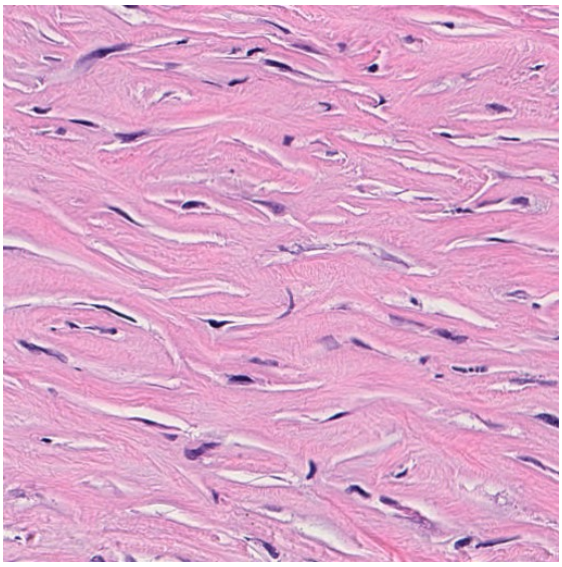

Figure 2.2: H&E staining of a tissue sample. Cell nuclei are stained dark blue to purple by hematoxylin, while cytoplasmic and extracellular components appear in shades of pink stained by eosin.

2.1.2 Structure

WSIs are extremely large digital files and high resolution images, often comprising millions of pixels. For instance, when scanned at 40x magnification, they can achieve a resolution of 0.25 micrometers per pixel. As a result, they require significant storage space, with uncompressed files occupying around 48 megabytes for every square millimeter of tissue [11]. To efficiently manage and navigate this vast amount of data, WSIs are structured in a multi-resolution pyramid format. At the top of the pyramid is the lowest resolution image, which is a thumbnail of the entire WSI. Each subsequent level down the pyramid represents a higher resolution version of the image, culminating in the highest resolution image at the base of the pyramid, where information like cellular details, and tissue structures are more easily discernible. This pyramidal structure is shown in Figure 2.3.

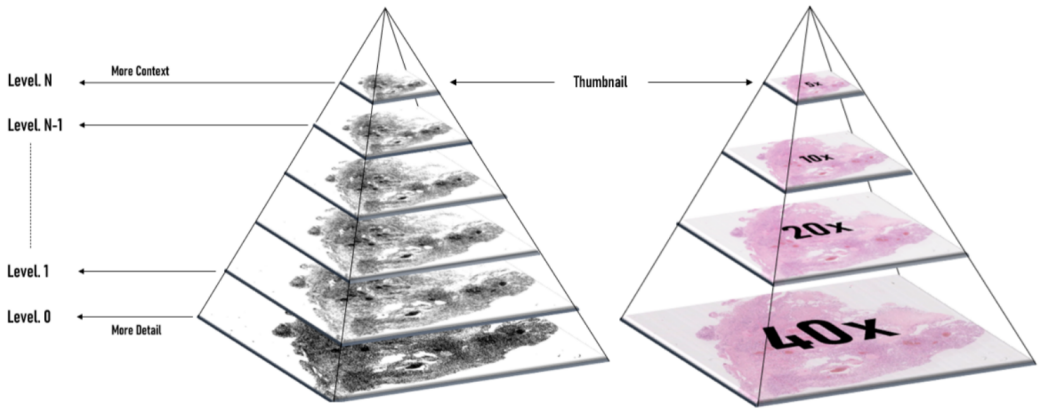

Figure 2.3: Structure of a WSIs in pyramid format, where it is possible to see the different levels of resolution. We have access to more details when we look at the lower levels of the pyramid [9].

This pyramidal structure allows for efficient data handling, as users can quickly access different levels of detail without needing to load the entire high-resolution image at once. It also facilitates faster image processing and analysis, as algorithms can operate on lower-resolution versions of the image when appropriate, reducing computational load and improving performance.

2.2 Pathological Overview

2.2.1 Anatomy of the Pleura

The pleura is a serous membrane formed from a layer of squamous mesothelial cells tightly attached by a network of dense connective tissue containing elastic and collagenous fibers (Figure 2.4). This serous membrane is composed of two components: the parietal pleura, which covers the internal surface of the thoracic cavity, including the chest wall and diaphragm; and the visceral pleura, which envelops the lung parenchyma, the pulmonary vessels, bronchi, and nerves [12].

The pleural space is a cavity delimited by the visceral pleura, which covers the lung, and the parietal pleura, which covers the chest wall, diaphragm, and mediastinum. This cavity contains a small amount of pleural fluid, estimated at 0.26 ml/kg of body weight, which is an ultrafiltrate of plasma mainly formed by filtration from the systemic capillaries of the parietal pleura. Its turnover is essential to ensure lubrication and mechanical connection between the lungs and the chest wall.

The pleural space plays an important role in respiratory mechanics, yet its exact physiological function in humans remains a debated topic. Among the various theories, the most widely supported one is that the pleural membranes act as an elastic interface between the lungs and the thoracic cage, thus allowing the lungs to undergo continuous shape changes while remaining mechanically connected to the rib cage.

Additionally, another proposed function concerns the pressure within the pleural space. At functional residual capacity, the pleural pressure is negative relative to atmospheric pressure, and this is thought to play a key role in maintaining lung inflation. Specifically, this negative pressure opposes the natural elastic recoil of the lung tissue, which would otherwise cause the lungs to collapse [13, 14].

2.2.2 Pleuritis

The inflammation of the parietal pleura results in what is defined as pleuritis (pleurisy), which is a clinical syndrome characterized by pleuritic chest pain due to the soreness of the intercostal and phrenic nerves [15]. The resulting pain is typically acute, sharp, and well localized, which demonstrates the high density of nociceptive fibers within the parietal pleura, and is generally worsened by respiratory movements that increase pleural friction, including, for example, deep inspiration, coughing, sneezing, or laughing [16]. Contrary, the visceral pleura inflammation is often asymptomatic because of its autonomic innervation. Most pleuritis cases are idiopathic conditions, nevertheless, there are cases in which it represents a secondary manifestation of an underlying pathological process involving

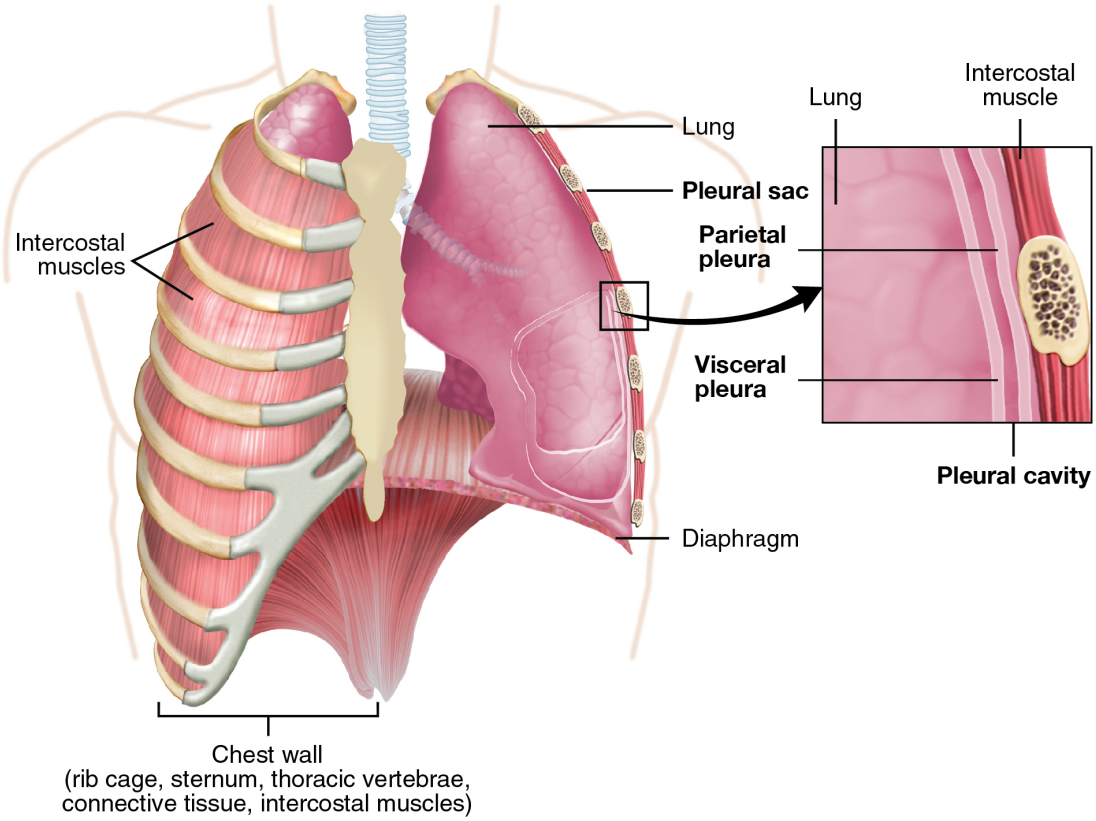

Figure 2.4: Anatomy of the pleura, showing the parietal pleura covering the chest wall and the visceral pleura enveloping the lung parenchyma, with a zoom-in on the pleural cavity.

the pleura, lung parenchyma, or nearby thoracic structures [16]. The predominant etiology is the infection one, with viral infections, particularly those caused by coxsackieviruses, influenza viruses, and respiratory syncytial virus, constituting the most frequent causes. Meanwhile, bacterial, mycobacterial, and fungal infections may also lead to pleural inflammation, especially in immunocompromised patients or in regions with high endemic prevalence [16]. These premises suggest that the correct diagnosis of chest pain represents a key aspect, because it can lead to life-threatening conditions, such as untreated inflammation of the parietal pleura. Hence, pleuritis requires a comprehensive and systematic diagnostic evaluation; early clinical analysis must prioritize the exclusion of critical etiologies, including pulmonary embolism, pneumonia with pleural involvement, pneumothorax, myocardial ischemia, pericarditis, and acute aortic pathology, before symptoms are attributed to benign causes. The diagnostic complexity of pleurisy is emphasized by evidence from surgical pleural biopsy studies, which demonstrate that

neoplastic diseases constitute the majority of cases (56%), with malignant pleural mesothelioma (23%), lung cancer (16%), and lymphoma (2.5%), which are the most frequently identified malignancies [17]. While infectious etiologies represent approximately 24% of the cases, with tuberculosis emerging as the leading cause (16.2%), followed by parapneumonic pleural effusion (3.6%), empyema (3.5%), non tuberculous mycobacterial infections (0.5%), and paragonimiasis (0.1%) [18]. These statistics are consistent with recent studies indicating that the pleura and lymph nodes represent the most involved extra pulmonary sites in tuberculosis [19, 20]. Moreover, noninfectious inflammatory conditions also contribute to the differential diagnosis: autoimmune diseases account for approximately 2.8% of pleural biopsy findings; among these, rheumatoid arthritis (1.3%) and systemic lupus erythematosus (0.3%) are the most prevalent [18].

In conclusion, these data emphasize that pleurisy should not be considered as a definitive diagnosis. Instead, it should be identified as a clinical manifestation that requires supplemental etiological clarification, which would help to define appropriate management strategies and prognostic evaluation [15].

2.2.3 Sarcomatoid mesothelioma

Mesothelioma is a malignant neoplasm of the pleura, originated by pleural mesothelial cells. This cell type plays a significant role in detecting and responding to noxious stimuli. The disease is strongly associated with long-term exposure to asbestos, a naturally occurring mineral fiber [21]. Early clinical manifestations are often vague and nonspecific, with symptoms including chest pain, dyspnea, fatigue and weight loss. Hence, initial misdiagnosis is very common, often mistaken for condition such as pneumonia [22]. In addition, tumor markers are not regularly expressed, and cancer typically demonstrates rapid progression. These factors contribute to further delayed diagnosis and an overall poor prognosis [21]. Histologically, mesothelioma is classified into three main subtypes: epithelioid, non-epithelioid, and sarcomatoid. Depending on the subclass, the prognostic outcomes differ; epithelioid mesothelioma is associated with the most favorable prognosis, followed by non-epithelioid, while sarcomatoid mesothelioma has the poorest prognosis. This is also due to the diagnostic challenge that sarcomatoid mesothelioma presents; indeed, its inconsistent expression of commonly used markers such as cytokeratin and calretinin turns the diagnostic evaluation complicated even though most mesothelioma can be diagnosed with biopsy and immunohistochemical staining [22]. Histologically, sarcomatoid mesothelioma is characterized by the presence of less than 10% epithelial tissue in biopsy specimens, along with spindle cells showing marked nuclear atypia [23, 24]. The differential diagnosis includes multiple malignant conditions: primary sarcomas of the chest wall, lung, abdominal wall, or peritoneum with pleural extension and solitary fibrous tumor of the pleura.

Consequently, distinguishing primary pleural tumors from secondary pleural involvement is a critical diagnostic step [22]. Moreover, sarcomatoid mesothelioma could be associated with coarse pleural calcification [22]. Therefore, it is crucial for radiologists to recognize this subtype in order to include it in the differential diagnosis of malignant pleural calcification and to avoid confusing it with benign calcified pleural plaque.

2.2.4 Differentiate between diseases

Differentiating benign pleural inflammation, such as pleuritis, from malignant pleural neoplasms, including rare cases such as sarcomatoid mesothelioma, remains a major diagnostic challenge due to overlapping clinical, radiologic, and pathologic features; indeed, both conditions may present with pleuritic chest pain, dyspnea, and pleural effusions, limiting the diagnostic value of clinical presentation alone [15, 21, 22]. Moreover, both conditions can present with spindle-cell proliferation and dense fibrous tissue, particularly in small biopsy specimens, making early recognition of malignant features difficult [25]. Therefore, the selection of appropriate diagnostic tools is crucial; although immunohistochemical markers can be useful, their irregular expression in sarcomatoid mesothelioma further complicates the differential diagnosis [25]. Additional techniques, such as p16/CDKN2A Fluorescence In Situ Hybridization (FISH), have proven valuable, as homozygous p16 deletion strongly favors a diagnosis of sarcomatoid mesothelioma over fibrous pleuritis [25]. Moreover, case reports have underlined instances where inflammatory pleuritis mimicked early sarcomatoid mesothelioma both macroscopically and microscopically, emphasizing the risk of misdiagnosis in the absence of comprehensive evaluation [17]. Within this context, recent findings in AI, particularly ML models, have shown promising potential to support the differentiation of benign and malignant pleural conditions; for instance, convolutional neural networks applied to thoracoscopic images can classify pleural lesions with high accuracy [26]. Furthermore, recent studies have shown that machine learning classifiers applied to DNA methylation profiles can distinguish pleural mesothelioma from chronic pleuritis with high accuracy. Additionally, deep convolutional neural networks trained on imaging data, such as FDG-PET/CT, have demonstrated good sensitivity and specificity [27, 28]. These emerging tools define the importance of implementing new approaches, capable of detecting subtle patterns beyond human perception. Particularly, these techniques allow the increasing of histopathologic and radiologic evaluation and potentially reducing observer variability, facilitating earlier and more accurate diagnosis of sarcomatoid mesothelioma.

2.3 Artificial Intelligence

When talking about AI, we refer to that branch of computer science that deals with producing systems capable of mimicking certain human cognitive abilities and performing tasks that normally require intellectual capacity typical of human beings. Some of these abilities include reasoning, learning complex patterns, understanding language, and the ability to recognize objects and images. These tasks are often difficult to define precisely with explicit and hard-coded rules, and this is why they were thought to be unsolvable through the use of traditional algorithms. In recent years, however, due to the increase in computational capacity with more powerful GPUs and TPUs, and the easier availability of enormous amounts of data, AI has made significant strides forward, leading to the development of increasingly sophisticated models in various fields, including medicine and digital pathology, among others.

The two main subtypes of AI are ML and Deep Learning (DL).

ML is a subset of AI that focuses on developing algorithms based on statistical models [29]. These algorithms include techniques such as Support Vector Machines (SVMs), Decision Tree, Random Forest, and regression models. Being statistical models, these algorithms learn parameters within shallow or predefined representations, rather than complex hierarchical ones. For this reason, they often require feature engineering, which is the manual selection and transformation of relevant features from raw data before model training. This process is extremely time-consuming and resource-intensive because it generally requires the intervention of domain experts to identify the most informative features.

ML, on the other hand, is a subset of ML that uses Deep Neural Networks (DNNs), composed of many layers of interconnected nodes, inspired by the structure of the human brain [30]. Each layer of the network learns increasingly abstract and complex representations of the input data, enabling the model to capture intricate patterns and relationships within the data without the need for manual feature engineering. DNNs are particularly effective in processing unstructured data, such as images, audio, and text, and have led to significant improvements in various fields, including computer vision, speech recognition, and natural language processing. The most commonly used type of ML models in computer vision are Convolutional Neural Networks (CNNs), due to their ability to learn hierarchical representations of images through the use of convolutional filters that capture spatial features at different levels of abstraction [31]. By applying learnable filters across an input image, CNNs are able to automatically extract relevant features such as edges, textures, and shapes in the initial layers, and more complex features such as objects and scenes in the deeper layers of the network. For this reason, CNNs are well-suited for image-related tasks such as classification, detection, and segmentation. Indeed, while not without limitations and having been surpassed in

some areas by more recent models like transformers, CNNs remain one of the most effective and widely used approaches for medical image analysis, including digital pathology. Pre-trained CNN architectures on large generic image datasets, such as ImageNet [32], are often adapted and fine-tuned on specific medical images, one of the most famous examples being the ResNet architecture [33].

2.3.1 AI in histopathology

Contrary to what one might expect, the medical field was among the first to adopt digital image processing and computerized image analysis, with early applications focused on the analysis of cells and microscopy images [34]. Already in 1965, M Mendelsohn et al. proposed a method for the use of automated image analysis [35, 36]. Individual cells in blood smears could be classified into subtypes based on quantitative cellular characteristics such as size, shape, and distribution of chromatin, in order to analyze blood composition and help diagnose a series of diseases. Initially, these methods were based on traditional image processing techniques and manual feature engineering, which made them limited in their ability to generalize to new data [37].

However, there was a great need to analyze images more effectively and to provide stronger support to pathologists, given that much time was spent on manual image analysis. For example, 80% of the 1 million prostate biopsies every year in the United States are benign, which means that 80% of the pathologist's time is spent analyzing healthy tissues [38].

The biggest turning point came in the early 2000s with the advent of WSIs, which allowed for the digitization of entire histopathological slides at high resolution, enabling data to be easily stored and shared. This, coupled with the increased computational power of modern hardware, was the perfect storm that allowed the application of AI techniques to histopathological image analysis.

Building on these advances, the past decade saw the first uses of DNNs, in particular CNNs, that departed from the use of manual feature engineering, which required domain knowledge in order to obtain good results [39]. ML, instead, allows learning hierarchical representations directly from data, learning relevant features automatically and proving easier and faster to implement while also achieving better performance. Naturally, the field of digital pathology is in continuous evolution, and new methods and architectures are proposed continuously, although some areas are less developed than others, mainly due to the scarcity of publicly available annotated data. This is exactly the case with the classification between pleuritis and sarcomatoid mesothelioma.

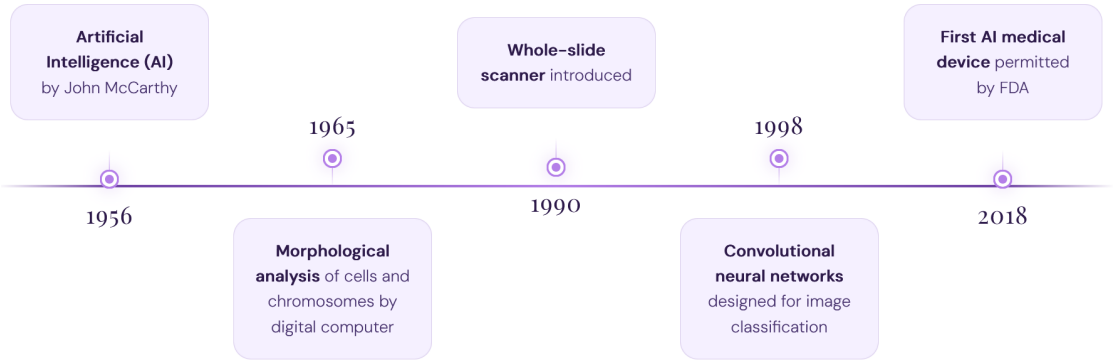

Figure 2.5: Advancements in digital pathology over the years.

2.3.2 Supervised Learning Models

Supervised learning is a subtype of ML in which the model is trained on a labeled dataset, that is, a set of input data each associated with a corresponding label or desired output value. In particular, in the case of WSIs, labels can be associated at both the whole image level (slide-level) and at the individual patch level. In the first case, the entire WSI is labeled with a single class that represents the overall diagnosis or pathological condition present in the sample. In the second case, instead, each patch extracted from the WSI is labeled individually, generally by domain experts such as pathologists as shown in Figure 2.6.

This approach allows the model to learn the relationships between the tissue and the provided labels, reducing the amount of data necessary for training compared to other approaches. However, having accurate patch-level annotation requires significant work from domain experts, and is usually extremely labor-intensive and costly. For this reason, this type of training is not always practicable, especially when dataset sizes are large. Regarding the specific problem of classification between pleuritis and sarcomatoid mesothelioma, there is only one study that addresses this topic from Julia R. Naso et al. and it uses a patch-level supervised learning approach [2]. The network presented in the paper is called SpindleMesoNet and is based on a ResNet-18 architecture pre-trained on ImageNet followed by a RNN layer to capture spatial dependencies between patches. The architecture was trained on a patch-level annotated by expert pathologist dataset of 58 malignant mesothelioma WSIs and 81 benign pleuritis. The model evaluation was performed on three distinct test sets, a training/cross-validation set with 5 fold cross-validation, an independent referral test set composed of 40 difficult cases submitted for expert consultation and an externally stained set with 39 cases from other institutions.

The model achieved an area under the ROC curve (AUC) of 93.2% on the training/cross-validation set, 92.5% on the referral test set, and 98.9% on the external set, indicating strong performances even on unseen data from different sources. The fundamental limitation of this approach reside in the need for annotated patch-level data. Even if the performance of this approach is good, it is difficult to apply it in real-world scenarios where even if WSIs are available, patch-level annotations are usually not.

This is why the following model types were implemented, in order to rely less on patch level annotations and only on slide-level labels if any.

2.3.3 Weakly supervised Learning Models

Although supervised learning has shown excellent performance, even in areas beyond the classification between pleuritis and sarcomatoid mesothelioma, the need for patch-level annotations represents a significant limitation for the applicability of these models in real-world scenarios. For this reason, there arose a need to develop models that could be trained using only slide-level labels, which are considerably easier to obtain. This type of approach falls within the Weakly-Supervised Learning (WSL) paradigm, which distinguishes itself in that annotations are typically incomplete, inexact, or inaccurate. In the context of WSI classification, annotations are most commonly incomplete, since only slide-level labels are available. This paradigm is commonly implemented through the MIL framework, in which each WSI is considered as a bag of unlabeled patches [40].

These models consist of two main phases, a feature extraction phase and a feature aggregation phase. In the first phase, patches are extracted from the WSI and passed through a neural network to extract a feature vector for each patch. Generally, in this phase a CNN pre-trained on a large dataset of generic images is used, for example ImageNet.

In the second phase, the extracted feature vectors are aggregated to produce a single prediction at the slide level. This can be done in several ways, for example by using a weighted average of the features or by using an attention mechanism to weight the importance of each patch, allowing the MIL model to automatically learn which regions contribute most to the final decision. Notable examples of such architectures include CLAM [41], which leverages attention-based pooling with instance-level clustering constraints; TransMIL [42], which incorporates a Transformer-based aggregator to capture morphological and spatial relationships among patches; and RRT-MIL [43], which introduces a re-embedding approach to improve the quality of patch representations before aggregation. These models will be described in greater detail in Chapter 5.

Despite the advantages in terms of scalability, MIL models remain sensitive to the quality of patch representation and can fail when diagnostic regions are

extremely rare or poorly distinguishable at the local level. They also require a relatively high number of WSIs to effectively learn the discriminative features between classes. Unfortunately, to the best of our knowledge, there are no studies that apply MIL models to the specific problem of classifying between pleuritis and sarcomatoid mesothelioma.

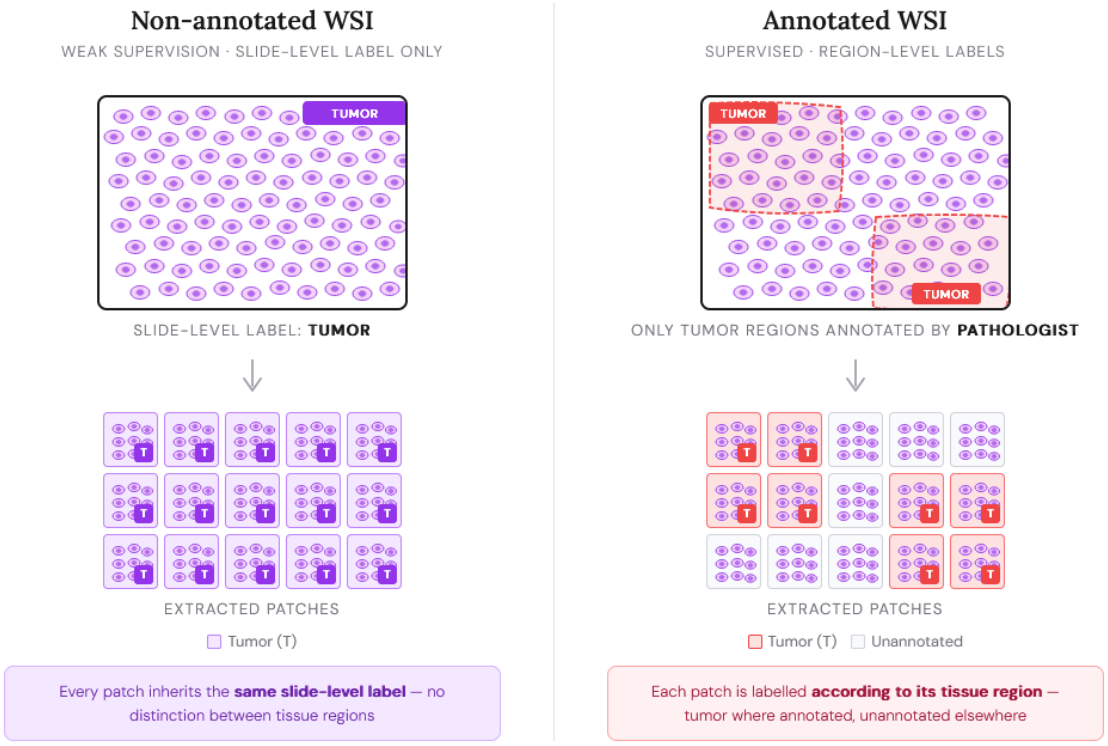

Figure 2.6: Difference between patch-level and slide-level annotations.

2.3.4 Self-supervised Learning Models

Self-Supervised Learning (SSL) improves on the previous WSL paradigm, as it does not require any labels, not even at the WSI level, for model training [44]. In this way, the need for annotations provided by domain experts is completely eliminated, making the model training more convenient and applicable to larger unlabeled datasets. SSL is based on the idea of creating pretext tasks that allow the model to learn useful representations from raw data without explicit supervision. These pretext tasks can include operations such as predicting missing parts in an image, predicting the rotation or relative position of patches extracted from an image, or even distinguishing between patches extracted from the same image and patches from different images. All these tasks push the model to capture relevant

structural and semantic features. These models are then usually used as feature extractors, and should have a deeper understanding of the tissue structure and relevant morphological patterns. The feature vectors extracted from the patches are then aggregated at the WSI level using techniques similar to those described in the previous paragraph on WSL, to produce a final prediction.

Although there are no specific studies applying SSL to this specific classification problem, many SSL models are used as foundation models in the field of digital pathology, being pre-trained on large datasets of unlabeled histological images. This, in theory, allows these models to learn rich and generalizable representations that can be effectively transferred to various downstream tasks. Among these, the most notable at the moment are UNI [45], with its subsequent version UNI2 [45], and CONCH [46].

UNI is a vision-only model designed to be used as a universal feature extractor for histopathological images, based on Vision Transformer (ViT) architecture, and uses the DINOv2 SSL framework [47]. The advantage of UNI is that it was trained on a massive dataset, one of the largest ever used in digital pathology, comprising over 100 million patches extracted from more than 100,000 WSIs from various organs. This variety allows the model to learn different types of tissue structures and morphological patterns, making it highly generalizable to different histopathological tasks. This has been demonstrated by the model across 34 different tasks, including classification, segmentation, and survival analysis, outperforming previous state-of-the-art models such as CTransPath [48] and REMEDIS [49] in many of them. The model is not without limitations, as it requires significant computational resources for both training and inference, and the architecture lacks vision-specific biases [45].

UNI's evolution is UNI2, which builds upon the great foundations of its predecessor, but introduces several improvements to enhance performance and efficiency. This model is still based on the ViT architecture, UNI2 has been trained on an even larger and more diverse dataset. The comparative analysis between UNI and UNI2 demonstrates that UNI2 outperforms its predecessor across various histopathological tasks, producing even more informative feature representations, and achieving more robustness to domain shifts. The model performs well even if tested on images from organs not present in the training dataset, and from different staining protocols and scanners [45].

A different approach is taken by CONCH, which is based on a hybrid architecture that combines CNNs and ViTs. Utilizing the framework CoCa (Contrastive Captioners) [50], CONCH also learns semantic relationships between image regions and their corresponding textual descriptions. Thanks to training on a large dataset of image-caption pairs, CONCH learns a multimodal representation that connects histological images with clinical language [46]. This allows the model to perform tasks that unimodal models cannot, such as zero-shot learning, where images can be

classified or retrieved using only text queries In Figure 2.7 the differences between supervised learning, WSL and self supervised Learning are summarized.

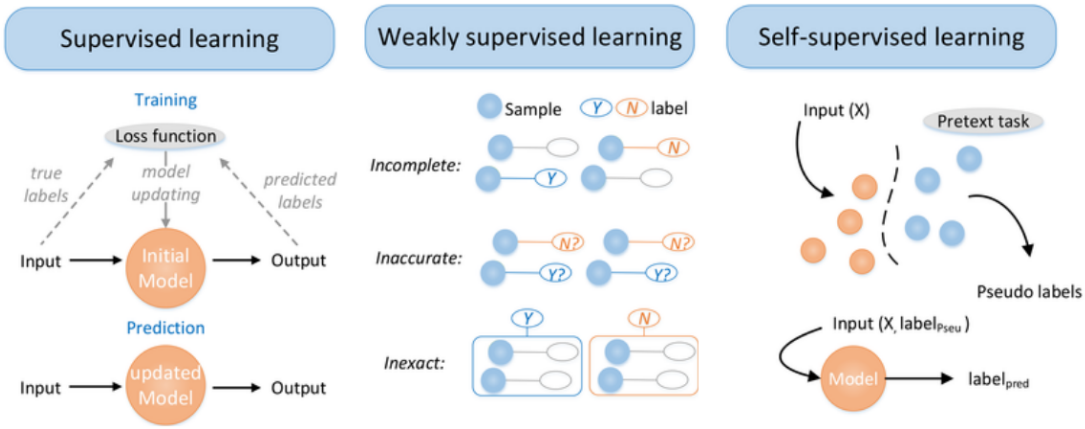

Figure 2.7: Differences between supervised learning, WSL and self supervised Learning [51].

2.4 Software, programming languages and libraries

There are a variety of tools that can be used for visualization, editing and annotation of WSIs, some of the most popular include Aperio ImageScope, QuPath [52], and Automated Slide Analysis Platform (ASAP) among others [53]. These software provide functionalities for viewing high-resolution images at different magnifications, annotating regions of interest, and performing basic image analysis tasks. Each tool has its own strengths and weaknesses, and the choice of which to use often depends on the specific requirements of the project and user preferences.

For this thesis, we decided to use ASAP, as it is an open-source software that is specifically designed for the visualization, annotation and basic analysis of WSIs. A notable advantage was its compatibility with various WSI formats, including ṅdpi files, which was the format of our dataset. It was also relatively easy to use and provided a range of useful features for annotating and analyzing histopathological images. In our use case the images were annotated by expert pathologists using Aiforia, a cloud-based platform for digital pathology that allows for collaborative annotation and analysis of WSIs. After that they were manually copied using ASAP, particularly useful was the rich set of annotation and drawing tools provided by the software, that were used to delineate regions of interest on the WSIs. These annotations are then stored in a simple, human-readable XML format, making it

easy to parse and use for further analysis. Additionally, the software provides a long list of real-time image processing and analysis tools, including color deconvolution, nuclei detection, and area measurements, which can be useful for basic image analysis tasks.

Underlying ASAP's functionalities is the OpenSlide library, a C library that is at the base of many digital pathology tools [54]. OpenSlide provides a simple and efficient way to read and manipulate WSIs in various formats, including Aperio, Hamamatsu, Leica, and others. In addition to its native C implementation, OpenSlide also has bindings for several programming languages, including Java, Ruby and Python, the latter was extensively used in this thesis in order to preprocess the WSIs and extract patches from them.

Figure 2.8 shows the ASAP software interface, displaying a WSI and some of the annotation tools available, like the possibility to create groups for different annotation regions, and the drawing tools, to delineate regions of interest.

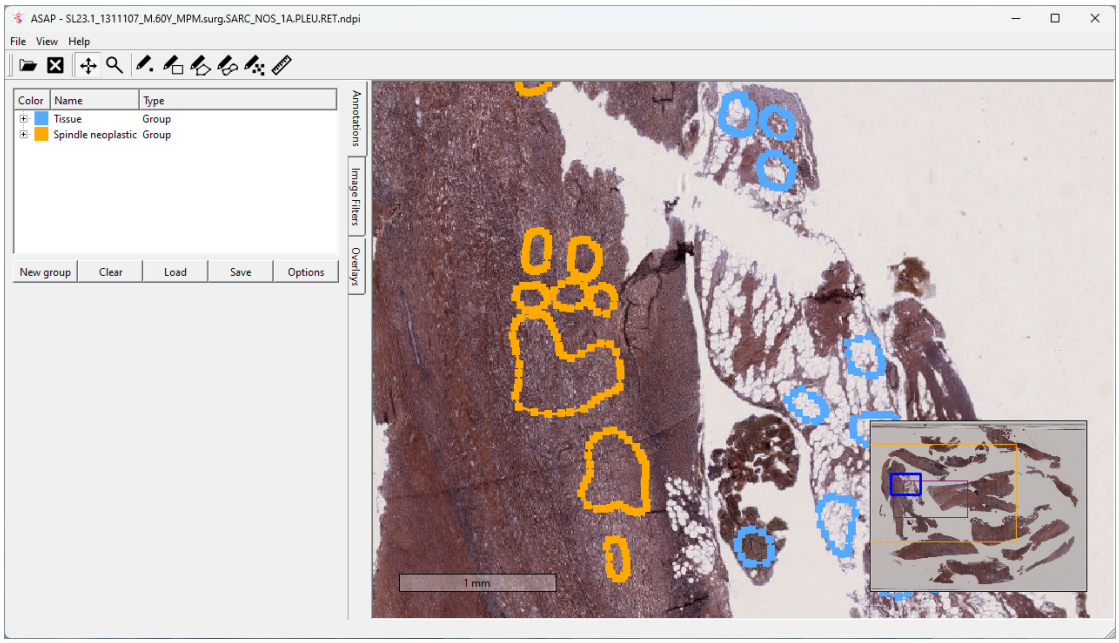

Figure 2.8: ASAP software interface showing a WSI and annotation tools.

For the implementation of the ML models, the Python programming language was used, due to its popularity in the data science and ML communities, as well as for its OpenSlide availability.

PyTorch was the main ML library used for the implementation of the models, due to its flexibility, ease of use, and community support in the research field especially [55].

Several other libraries were used for various tasks, including NumPy [56] and Pandas for data manipulation, Matplotlib for data visualization, and Scikit-learn [57] for evaluation metrics and other ML utilities.

Chapter 3

Dataset

In this work, the dataset that was used to train the feature aggregators is not publicly available online, since it was provided by the University of Microbiology of Turin, and the data is strictly confidential due to privacy reasons. Unfortunately there is no publicly available dataset at the moment of writing this that enables us to perform classification on sarcomatoid mesothelioma and pleuritis, for this reason and the limited number of data available to us, we were not able to perform testing on a complete external dataset, settling with some sarcomatoid WSIs that we were able to find online, from the University of Leeds Virtual Pathology Project Website [58].

3.1 Dataset composition

The dataset is composed by 54 WSIs of pleural biopsies, stained with H&E, from two different hospitals in Turin: San Luigi and Molinette. They were acquired in two different years, 2023 and 2025, and were digitalized at different maximum magnification: 40x and 20x respectively. The WSIs are divided in two classes: 21 images are sarcomatoid mesothelioma and 33 are pleuritis. The mesothelioma sarcomatoid cases were all obtained in the year 2023, with the exception of one case from 2025, whereas the pleuritis cases are all from 2025. The composition of the dataset is summarized in Table 3.1.

It is important to note that using this dataset for training a ML model is extremely challenging, due to the limited number of samples available, the class imbalance and the fact that such an important part of the images acquisition process, the magnification level and and consequently the Microns Per Pixel (MPP) value, is different between the two classes.

Class	2023 (40x)	2025 (20x)	Total
Sarcomatoid Mesothelioma	20	1	21
Pleuritis	0	33	33
Total	20	34	54

Table 3.1: Distribution of WSIs per class and year of acquisition / magnification, it is evident how the dataset is imbalanced and how the magnification differs between classes.

3.1.1 External cohort

An external cohort was collected from the University of Leeds Virtual Pathology Project website [58]. Since this publicly available dataset is primarily focused on cancer cases, only sarcomatoid mesothelioma slides were available, resulting in a total of 5 whole slide images. It was the only public dataset we were able to find, as no publicly available examples of pleuritis, or other sarcomatoid cases could be found. Despite its very limited size, it was included in the evaluation because it represented the only source of external data and was considered more valuable than having no external validation at all.

3.2 Dataset annotations

Of these images only 6 of the H&E stained WSIs were annotated by an expert pathologist, marking regions that contained tumorous tissue or pleuritis on the online platform Aiforia. The annotations were also unbalanced per class, with 5 images of pleuritis and only 1 of sarcomatoid mesothelioma.

In order to take advantage of these annotations, the Region of Interests (ROIs) were manually copied from Aiforia using ASAP, in this way we could use a more flexible XML format. The annotations were only used during the training of the supervised models, whereas the feature extraction models were never fine-tuned on the annotated data. The distribution of the annotations is summarized in Table 3.2.

Class	**H&E WSIs**	**H&E Annotated**
Sarcomatoid Mesothelioma	21	1
Pleuritis	33	5
Total	54	6

Table 3.2: Number of annotated H&E WSIs per class, showing the limited number of annotations available for training supervised models.

Chapter 4

Supervised Methodology

4.1 Data Preprocessing

The data preprocessing pipeline follows the approach described in Naso et al. [2] and consists of two sequential steps: tissue segmentation, in which a binary mask of the tissue regions is produced, and patch extraction, in which image patches are cropped from the masked regions and saved to disk.

4.1.1 Tissue Segmentation

A low-resolution thumbnail is obtained by downsampling each WSI by a factor of 16 (Figure 4.1a) and converted to grayscale (Figure 4.1b). A Laplacian filter is then applied to improve tissue-background separation by enhancing boundaries and suppressing uniform regions. (Figure 4.1c). The filtered image is binarised using Otsu's method [59], which automatically selects the threshold that maximises the separation between foreground and background intensity distributions (Figure 4.1d). Binary closing is then applied with a disk-shaped structuring element of radius 50 µm to fill small gaps and smooth the mask boundaries (Figure 4.1e). Finally, connected components smaller than 10^5 µm^2 are removed, discarding fragments too small to contain diagnostically relevant content, and residual small holes within tissue regions are filled using the same area threshold, yielding the final tissue mask (Figure 4.1f). If the tissue percentage resulting from the Laplacian step falls outside a plausible range (below 8% or above 70%), the pipeline retries without the Laplacian filter, falling back to direct Otsu thresholding on the grayscale image. The resulting tissue mask is subsequently used to restrict patch extraction to tissue-containing regions.

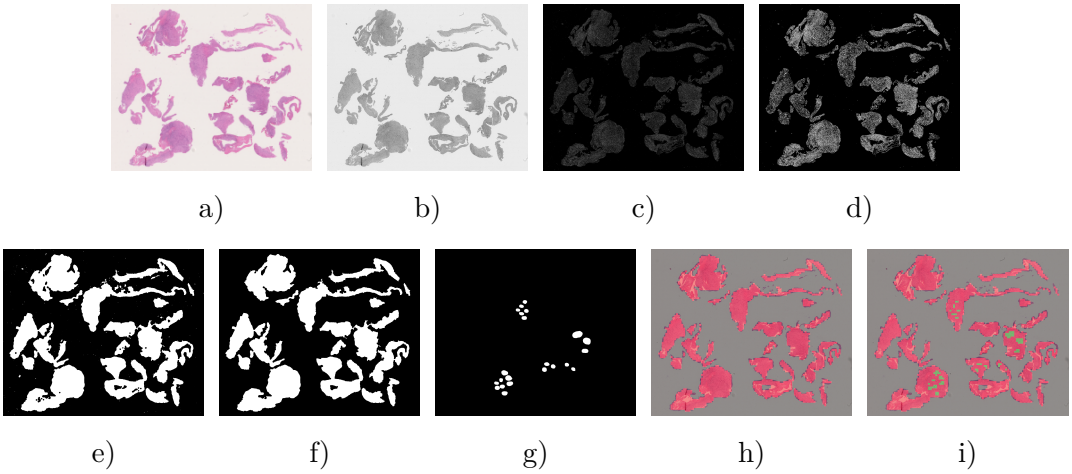

Figure 4.1: Intermediate outputs of the tissue segmentation pipeline. **a)** Downsampled RGB thumbnail. **b)** Grayscale conversion. **c)** Laplacian-filtered image. **d)** Binary mask after Otsu's thresholding. **e)** Mask after morphological closing. **f)** Final tissue mask after removal of small objects and holes. **g)** Pathologist annotation mask (training slides only). **h)** Extracted patches overlaid on the tissue mask. **i)** Extracted patches overlaid on the tissue mask in red the ones always extracted, and in green the ones extracted during training that correspond to the pathologist annotation.

4.1.2 Patch Extraction

Patches of 512×512 pixels are extracted using a sliding window with no overlap. Each candidate location is accepted only if at least 90% of its area falls within the tissue mask. All patches are extracted at a fixed physical resolution of 0.44 µm/pixel, rescaling the read region as needed based on the slide MPP, so that the patch size corresponds to a consistent tissue area across all slides. Unlike the weakly supervised pipeline, patches are read from the WSI immediately and saved to disk as PNG files.

For training slides, pathologist annotations are available as polygon contours in ASAP XML format, manually transcribed from region-level labels provided on the Aiforia platform (Figure 4.1g). The annotations identify two tissue categories: *Spindle neoplastic* (sarcomatoid mesothelioma) and *Spindle reactive* (pleuritis). A patch is included in the training set only if at least 75% of its area overlaps with an annotated polygon. This filtering is not applied to validation and test slides. The resulting patch distributions, with and without annotation filtering, are shown in Figures 4.1h–i.

4.2 Supervised Architecture

The supervised architecture also mimics the pipeline proposed by Naso et al. [2] and operates in two stages: a patch-level classifier is first trained on annotated patches, and its predictions are then aggregated across all patches of a slide to produce a slide-level score.

4.2.1 Patch-level Classifier

The patch-level classifier is a ResNet-18 [60] pretrained on ImageNet. Only the final fully connected layer is replaced with a two-class linear classifier and trained, while the convolutional backbone is kept frozen. This choice is motivated by the small size of the annotated training set: fine-tuning the full network would risk overfitting, whereas freezing the backbone allows the model to retain the general visual representations learned on ImageNet and adapt only the classification head to the histopathology domain.

4.2.2 Slide-level Aggregation

At inference time, the trained classifier is applied to all patches extracted from a WSI, producing a malignancy probability $p_k \in [0,1]$ for each patch k. Slide-level prediction follows the MIL-pool strategy of Naso et al. [2]: all patches are ranked by their malignancy probability, the top 0.5% are selected (with a minimum of 10 patches), and their probabilities are averaged to produce the final slide-level score. This approach assumes that a slide labeled as sarcomatoid mesothelioma contains at least some highly malignant patches, while a pleuritis slide contains none, consistent with the standard MIL assumption.

Chapter 5

Weakly Supervised Methodology

In this chapter, the approach adopted by the weakly supervised methodology is described. Starting from the data preprocessing phases, which differ from those used in the supervised method due to the absence of annotations, the pipeline follows the standard approach adopted by the CLAM [41] architecture and commonly used in the literature. The architectures employed in this approach are then presented, including the feature extractors and the aggregators.

5.1 Data preprocessing

Data preprocessing for the weakly supervised approach differs from the supervised one in a fundamental way: instead of extracting and storing the actual patch images, only the patch coordinates are saved. This choice is motivated by the scale of the data involved, since extracting and storing all patches from every WSI at full resolution would require too much storage. By saving only coordinates, patch images are read on-the-fly from the WSI during feature extraction and loaded in batches directly by the dataloader. The preprocessing pipeline consists of three stages:

- **Tissue segmentation:** a binary mask of the foreground tissue regions is computed for each WSI, identifying the areas from which patches will be sampled and explicitly excluding background, holes, and artifacts.
- **Patch extraction:** a uniform grid of candidate patch positions is generated within the segmented tissue regions. Each candidate is tested against the tissue contours using the four-point criterion, and only the accepted coordinates are

saved to an HDF5 file alongside the patch metadata (size, level, downsample factor). No pixel data is stored at this stage.

- **Patch stitching:** a downscaled overview image is generated by placing the accepted patches on a canvas at a lower resolution level, providing a lightweight visual verification of the segmentation and coordinate extraction results.

Each stage is described in detail below.

5.1.1 Tissue Segmentation

The goal of tissue segmentation is to automatically identify the regions of the slide that contain biologically relevant tissue, and to exclude the background, i.e., the empty white areas of the glass slide, as well as artifacts such as dust, air bubbles, or scanner defects. Since operating at full resolution would be computationally impossible, the segmentation is performed at a downsampled pyramid level, which reduces the slide to a manageable resolution. The contours identified during tissue segmentation are then rescaled back to the full-resolution coordinate for the next steps of patch extraction and patch stitching.

The segmentation pipeline proceeds through the following steps, whose intermediate outputs are illustrated in Figure 5.1.

Step 1 - Thumbnail creation The slide is read at `seg_level` using OpenSlide, producing a RGBA image (the alpha channel is subsequently discarded) downsampled of a factor of 64× in respect to the original slide, as shown in Figure 5.1a. Working at this reduced resolution allows the entire slide to fit in memory and speeds up the segmentation process.

Step 2 - HSV conversion and saturation extraction. The RGB image is converted to the HSV (Hue, Saturation, Value) color space, whose three channels are shown in Figure 5.1b-d. This choice is motivated by the physical properties of H&E staining, since tissue stained with H&E appears in shades of purple and pink, i.e., it is chromatically saturated, whereas the glass background is white and therefore has saturation close to zero. Among the three channels, the Saturation channel S Figure 5.1c is selected to be used in the next steps, since it provides the clearest separation between tissue and background.

Step 3 - Median blur. A median filter with kernel size 15×15 is applied to the Saturation channel before thresholding (Figure 5.1e). The filter replaces each pixel value with the median of its neighborhood, effectively suppressing isolated noise pixels, such as scanner artifacts or dust, which appear as abrupt local deviations from the surrounding tissue signal.

Step 4 - Binary thresholding. The blurred Saturation channel is binarized to produce a binary mask in which each pixel is classified as tissue (255) or background (0) (Figure 5.1f). Two strategies are available: a fixed threshold (`sthresh` = 15), suitable when the staining quality is consistent across slides, and Otsu's method [59], which automatically determines the optimal threshold by minimizing the intra-class variance of the pixel intensity distribution.

Step 5 - Morphological closing. A morphological closing operation (dilation followed by erosion) with a 4×4 kernel is applied to the binary mask (Figure 5.1g). Closing fills small holes that appear inside tissue regions due to lightly stained areas or staining artifacts, and merges nearby fragmented regions into a single compact contour. Without this step, the contour detection in the following stage could produce a large number of spurious micro-contours, increasing both memory usage and processing time.

Step 6 - Contour detection. Contours are extracted from the binary mask using `cv2.findContours` (Figure 5.1h). The retrieval mode `RETR_CCOMP` is adopted because it organizes the detected contours into a two-level hierarchy: outer boundaries correspond to foreground tissue regions, while inner boundaries correspond to holes enclosed within them, such as vascular lumina or tissue folds. This distinction is exploited directly in the subsequent filtering and patching stages, where holes must be tracked separately and excluded from patch extraction.

Step 7 - Contour filtering. The output of contour detection can contain contours corresponding to dust particles, scanner artifacts, or other irrelevant structures. Contours are therefore filtered by their effective area, defined as the contour area minus the sum of the areas of its enclosed holes. Only contours whose effective area exceeds a minimum threshold are retained as valid tissue regions (Figure 5.1i). Similarly, holes are kept only if their area exceeds a minimum threshold, and at most `max_n_holes` holes per contour are retained. Both thresholds are expressed as multiples of a reference patch area scaled to the segmentation level, making the criterion invariant to the choice of `seg_level`.

Step 8 - Contour rescaling. The contours extracted at `seg_level` are expressed in the coordinate system of that downsampled level and must therefore be converted to level-0 coordinates before they can be used for patch sampling. This is done by multiplying each contour point by the downsample factor of `seg_level`, i.e., the ratio between the full-resolution slide dimensions and the segmentation-level dimensions. The rescaled contours serve as vector masks in the level-0 coordinate system: during patch extraction, each candidate position is tested for inclusion

against these contours using `cv2.pointPolygonTest`, independently of the pyramid level at which the patch pixel data will eventually be read. The final segmentation overlay on the thumbnail extracted in Step 1 is shown in Figure 5.1j.

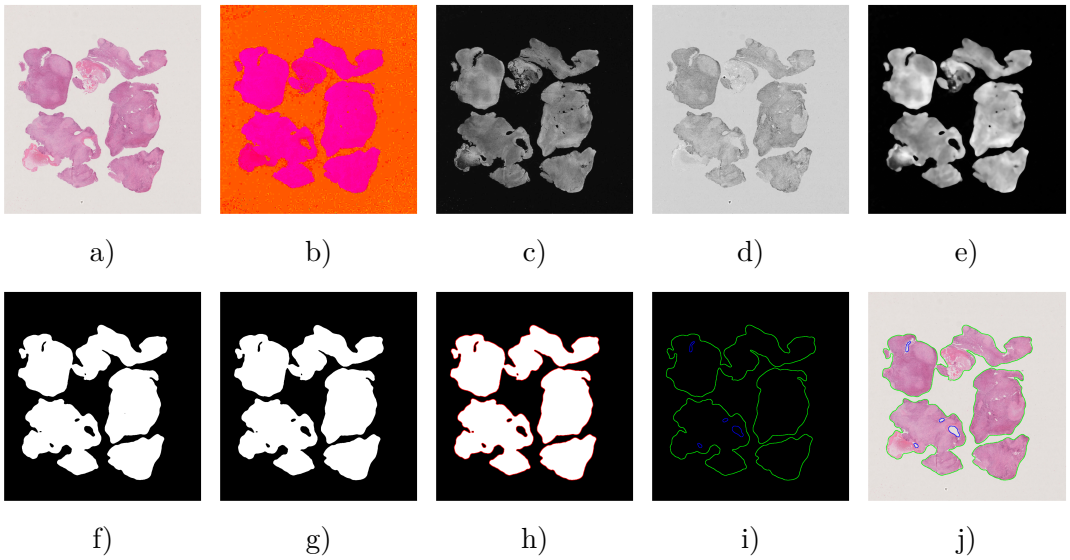

Figure 5.1: Intermediate outputs of the tissue segmentation pipeline for a representative H&E stained WSI. **a)** Original RGB image at `seg_level` (downsample factor $\approx 64\times$). **b)** Hue channel. **c)** Saturation channel: stained tissue appears bright while the glass background is dark (saturation ≈ 0). This channel is used in all subsequent steps. **d)** Value channel. **e)** Saturation channel after median blur (15×15 kernel). **f)** Binary mask after thresholding (`sthresh`$= 15$). **g)** Binary mask after morphological closing. **h)** Contours detection before filtering. **i)** Contours after filtering: tissue boundaries (green) and holes (blue). **j)** Final segmentation overlay: tissue boundaries (green) and holes (red) delimit the regions from which patches are extracted.

5.1.2 Patch Extraction

For each tissue contour, patch coordinates are generated by sliding a 512×512 pixel window across the contour's bounding box with a stride equal to the patch size, producing a regular grid of non-overlapping candidate coordinates. Each candidate is then tested for inclusion using the *four-point* criterion: a patch is accepted if all four of its inner corner points lie within the tissue contour and outside any of its associated holes, as determined by `cv2.pointPolygonTest`. The accepted coordinates are saved to an HDF5 file together with the patch metadata (size, level, and downsample factor), while the actual pixel content is not read at this stage.

Matching Physical Resolution Across Cohorts.

The dataset used in this work comprises slides from two different acquisition cohorts: slides acquired in 2023 at 40× nominal magnification, and slides acquired in 2025 at 20× nominal magnification. Although the nominal magnifications differ, it is the physical resolution expressed in MPP that determines the actual content of each patch: a lower MPP means that each pixel covers a smaller physical area, so biological structures such as cell nuclei appear larger in the image. If patches from the two cohorts were extracted at different MPP values, the same biological structure would occupy a different number of pixels depending on the cohort, making the feature representations incomparable and potentially introducing a systematic bias in the downstream models. It is therefore essential to ensure that all patches are extracted at the same MPP, regardless of the nominal magnification of the scanner used. Let MPP_ℓ denote the physical resolution at pyramid level ℓ, and let d_ℓ denote the downsample factor at that level relative to level 0. Then:

$$\mathrm{MPP}_\ell = \mathrm{MPP}_0 \cdot d_\ell \tag{5.1}$$

For the 2023 cohort, scanned at 40×, the base resolution is $\mathrm{MPP}_0^{(40\times)} \approx 0.22\ \mu\mathrm{m/px}$. At pyramid level 1, the downsample factor is $d_1 = 2$, giving:

$$\mathrm{MPP}_1^{(40\times)} = 0.22 \times 2 = 0.44\ \mu\mathrm{m/px} \tag{5.2}$$

For the 2025 cohort, scanned at 20×, the base resolution is already $\mathrm{MPP}_0^{(20\times)} \approx 0.44\ \mu\mathrm{m/px}$, so level 0 directly provides the target resolution:

$$\mathrm{MPP}_0^{(20\times)} = 0.44\ \mu\mathrm{m/px} \tag{5.3}$$

Since $\mathrm{MPP}_1^{(40\times)} = \mathrm{MPP}_0^{(20\times)} = 0.44\ \mu\mathrm{m/px}$, patches from the two cohorts represent tissue regions of identical physical extent:

$$512\ \mathrm{px} \times 0.44\ \frac{\mu\mathrm{m}}{\mathrm{px}} = 225\ \mu\mathrm{m} \quad \Rightarrow \quad \text{physical patch size} = 225 \times 225\ \mu\mathrm{m}^2 \tag{5.4}$$

To exploit this equivalence, patches from the 2023 cohort are extracted at `patch_level` = 1, while patches from the 2025 cohort are extracted at `patch_level` = 0.

5.1.3 Patch Stitching

To visually verify the segmentation and patch extraction results, a stitched overview image is generated for each slide by placing the accepted patches on a canvas at their downscaled positions (Figure 5.2). The stitching is performed by re-reading the WSI at a visualization level closest to a 64× downsample factor, producing a lightweight reconstruction of the tissue regions selected.

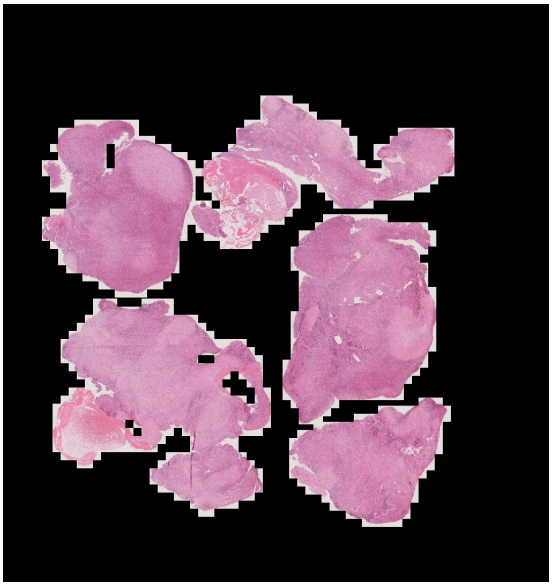

Figure 5.2: Stitched overview image for visual verification of the segmentation and patch extraction.

5.2 Weakly Supervised Architecture

The weakly supervised pipeline adopted in this work consists of two sequential stages: feature extraction, in which each patch is mapped to a compact feature vector by a pre-trained encoder, and feature aggregation, in which a MIL model combines the patch-level representations into a single slide-level prediction, as illustrated in Figure 5.3. Four encoders and three aggregators were evaluated in order to assess the impact of each component on classification performance.

5.2.1 Feature Extractors

Given the large number of patches per slide a feature extraction step is applied prior to MIL training to obtain a more compact representation for each patch. Each patch is passed through a pre-trained encoder operating in inference mode, producing a fixed-dimensional feature vector that summarizes the visual content of the patch. All patches were resized to 224×224 pixels prior to encoding, consistent with the input resolution expected by each model. The resulting feature vectors were saved in PyTorch `.pt` format and HDF5 files. Four encoder architectures were evaluated, differing in training strategy and pretraining domain: ResNet-50 is the only encoder not pretrained on histopathological data, and serves as a baseline against which the domain-specific encoders can be compared. Table 5.1 summarizes their key characteristics.

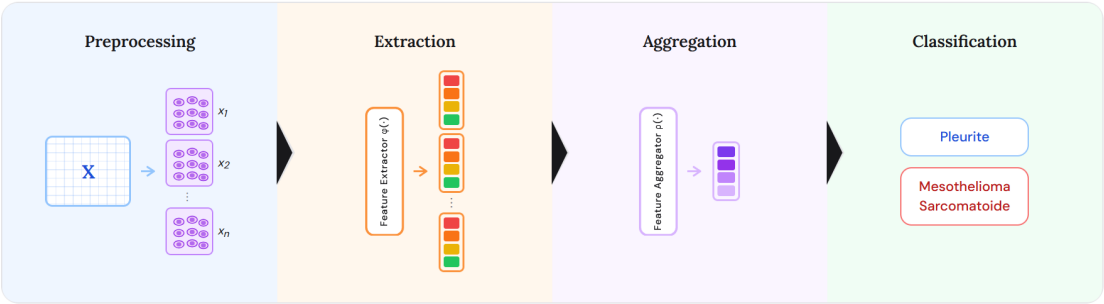

Figure 5.3: Overview of the weakly supervised MIL pipeline. After preprocessing each patch is encoded by a pre-trained feature extractor $\phi(\cdot)$ into a fixed-dimensional feature vector. The set of patch features forms a bag, which is processed by a feature aggregator $\rho(\cdot)$ to produce a slide-level representation, from which the final classification is made.

ResNet-50.

A ResNet-50 [60] pre-trained on ImageNet via supervised classification is included as a baseline. The truncated version of the network, without the final classification head, is used to produce a 1024-dimensional feature vector per patch. Although ResNet-50 was not designed for histopathology, it has been widely used as a feature extractor in MIL pipelines for computational pathology [41], and serves here as a reference against which the domain-specific encoders can be compared.

UNI v1.

UNI [45] is a self-supervised vision foundation model for computational pathology, based on a Vision Transformer Large (ViT-L/16) architecture with approximately 300 million parameters (Figure 5.4). It was pre-trained using the DINOv2 framework [47] on *Mass-100K*, a large proprietary dataset comprising over 100 million tissue patches sampled from 100,000 diagnostic H&E stained WSIs across 20 major tissue types, collected from Massachusetts General Hospital, Brigham and Women's Hospital, and the GTEx consortium. DINOv2 [47] is a self-supervised training framework based on self-distillation: a student network is trained to match the output of a teacher network, where the teacher is an exponential moving average of the student weights rather than a separately trained model. Both networks share the same architecture, and the student is trained on heavily augmented views of the input while the teacher sees less augmented versions, forcing the student to learn representations that are invariant to appearance changes. The training objective combines a cross-entropy loss between the student and teacher output distributions with a Masked Image Modeling (MIM) loss, in which a portion

of the input patches is masked and the student must reconstruct the teacher's representations for the missing regions. No labeled data is required at any stage. Evaluated across 34 computational pathology tasks, UNI demonstrated state-of-the-art performance compared to prior encoders. Using this encoder each patch is mapped to a 1024-dimensional feature vector.

UNI v2.

UNI v2 is the second generation of the UNI foundation model, released in January 2025, and follows the same DINOv2 training protocol described for UNI v1 (Figure 5.4). The main differences with respect to UNI v1 are the model scale and the pretraining dataset: UNI v2 adopts a ViT-H/14 architecture with 681 million parameters and was pretrained on over 200 million image tiles sampled from more than 350,000 WSIs, including both H&E and IHC slides sourced from Mass General Brigham. With this model each patch is encoded into a 1536-dimensional feature vector.

CONCH.

CONCH (CONtrastive learning from Captions for Histopathology) [46] is a vision-language foundation model for computational pathology. Unlike the UNI models, which rely solely on image-level self-supervised objectives, CONCH was pretrained by exploiting paired image-text supervision derived from histopathology reports and educational materials.

The pretraining dataset consists of 1.17 million human histopathology image-caption pairs assembled from Pub Med Central Open Access through an automated cleaning pipeline, illustrated in Figure 5.5a. Training follows the CoCa framework [50] and jointly optimizes two objectives: a *contrastive loss* that aligns image and text representations in a shared embedding space, and a *captioning loss* that trains a multi modal fusion decoder to generate captions conditioned on the image tokens. The combination of these two objectives encourages the image encoder to produce representations that are grounded in pathological language concepts, rather than purely visual ones.

In this work, CONCH is used exclusively as a feature extractor: captions and the text encoder are not involved at inference time, and only the image encoder is applied to extract a 512-dimensional feature vector from each patch.

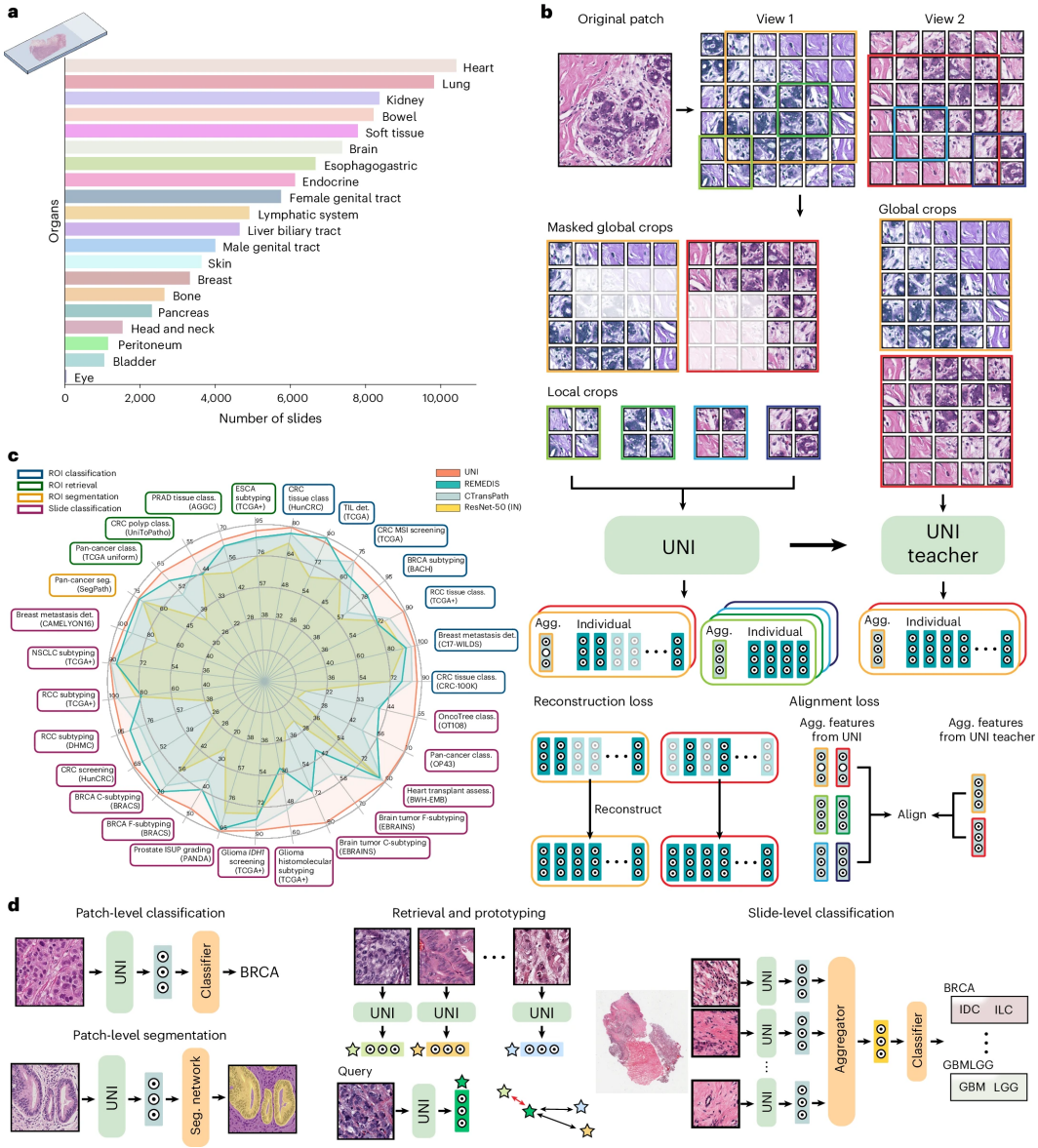

Figure 5.4: Overview of the UNI architecture adapted from [45]. **a)** Distribution of the Mass-100K dataset used for pretraining. **b)** UNI is pretrained on Mass-100K using the DINOv2 self-supervised framework, which combines MIM and self-distillation objectives. **c)** Performance comparison showing that UNI outperforms other pretrained encoders across 34 pathology-related tasks. **d)** Overview of the evaluation benchmark, including ROI classification, segmentation, image retrieval and prototyping, and slide-level classification tasks.

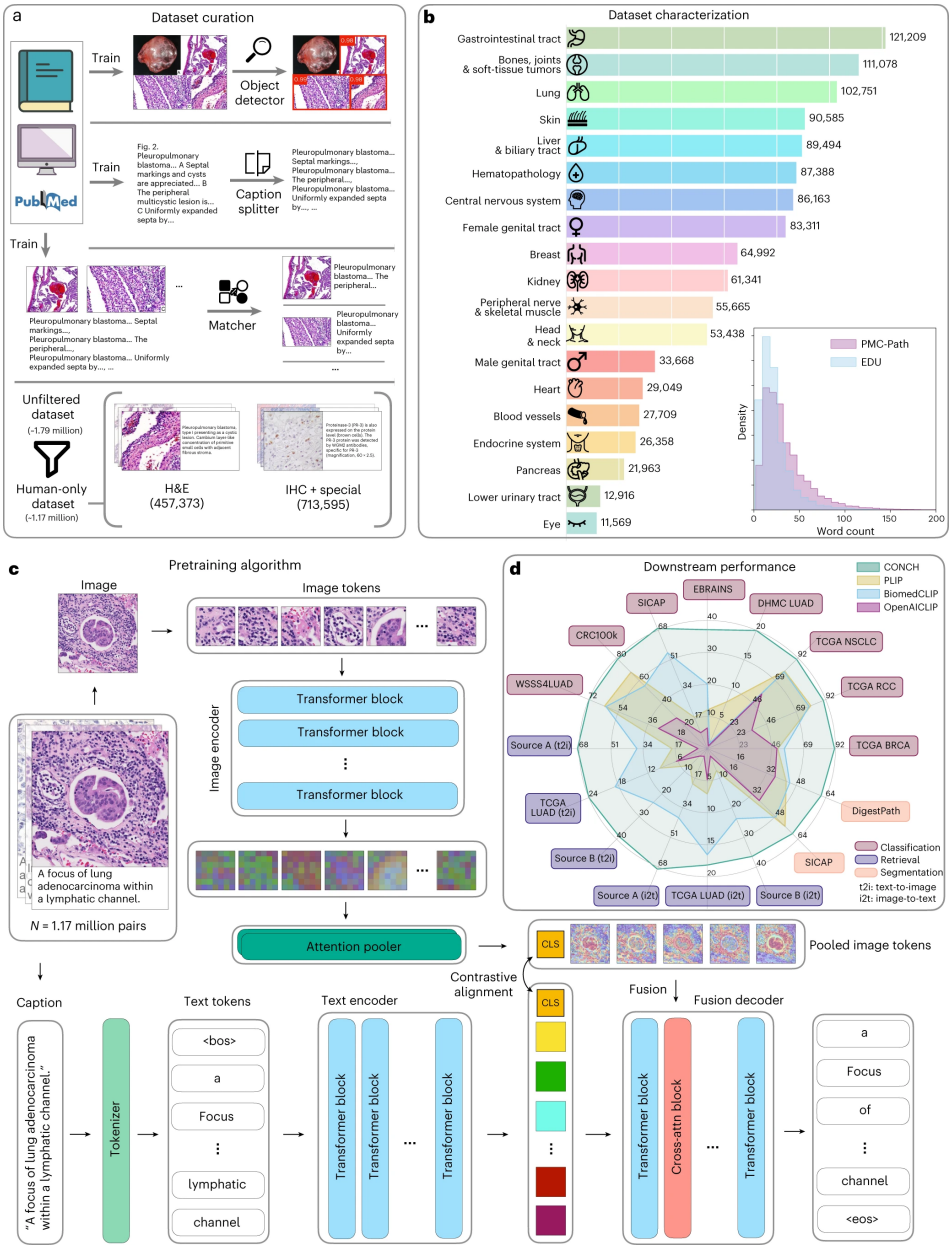

Figure 5.5: Overview of the CONCH pretraining pipeline, adapted from [46]. **a)** Automated data curation pipeline used to assemble the 1.17 million image-caption pairs from PubMed Central Open Access and educational sources. **b)** Distribution of the pretraining dataset across pathology topics. **c)** Pretraining architecture: an image encoder and a text encoder are **d)** Downstream performance comparison on zero-shot classification, retrieval, and segmentation tasks.

Encoder	Architecture	Paradigm	Pretraining data	Feat. dim.
ResNet-50	ResNet-50	Supervised	ImageNet	1024
UNI v1	ViT-L/16	SSL (DINOv2)	Mass-100K	1024
UNI v2	ViT-H/14	SSL (DINOv2)	200M Patches	1536
CONCH	ViT-B/16	VLP (CoCa)	1.17M Img-Text pairs	512

Table 5.1: Summary of the feature encoders evaluated in this thesis.

5.2.2 Feature Aggregators

Three MIL aggregation architectures were evaluated to classify each WSI as either sarcomatoid mesothelioma or pleuritis, based on the bag of patch-level features described in the previous section. All three models are trained end-to-end using the slide-level cross-entropy loss, and differ primarily in how they combine patch features into a slide-level representation.

CLAM

CLAM (Clustering-constrained Attention Multiple Instance Learning) [41] is an attention-based MIL framework specifically designed for computational pathology. Its architecture, illustrated in Figure 5.6, processes each bag through three sequential components. First, a shared fully-connected layer maps each patch feature $\mathbf{z}_k$ from its original dimensionality to a 512-dimensional embedding $\mathbf{h}_k$. Second, a gated attention module computes a scalar weight a_k for each patch. The gating mechanism combines two parallel linear projections of $\mathbf{h}_k$: one passed through a tanh activation and one through a sigmoid σ, which are multiplied element-wise and then linearly combined:

$$a_k = \frac{\exp\left\{\mathbf{w}^\top \left(\tanh(\mathbf{V}\mathbf{h}_k) \odot \sigma(\mathbf{U}\mathbf{h}_k)\right)\right\}}{\sum_j \exp\left\{\mathbf{w}^\top \left(\tanh(\mathbf{V}\mathbf{h}_j) \odot \sigma(\mathbf{U}\mathbf{h}_j)\right)\right\}} \tag{5.5}$$

where $\mathbf{w}$, $\mathbf{V}$ and $\mathbf{U}$ are learnable parameters. The sigmoid gate $\sigma(\mathbf{U}\mathbf{h}_k)$ acts as a soft mask that suppresses or amplifies each dimension of the tanh projection independently, allowing the model to selectively attend to specific feature dimensions rather than treating all dimensions uniformly. The slide-level representation is then computed as the attention-weighted sum $\mathbf{M} = \sum_k a_k \mathbf{h}_k$, which is passed to a linear classifier to produce the final prediction via a standard cross-entropy loss $\mathcal{L}_{\text{slide}}$.

Beyond standard attention pooling, CLAM introduces an *instance-level clustering loss* $\mathcal{L}_{\text{inst}}$. At each training step, the top-k and bottom-k (with $k = 8$) attended patches are selected and assigned pseudo-labels as positive or negative instances for that class, respectively. A per-class instance classifier is then trained on these pseudo-labeled patches using a cross-entropy loss, encouraging the attention module

to concentrate on diagnostically relevant regions. The total training objective combines the two losses as:

$$\mathcal{L} = 0.7 \cdot \mathcal{L}_{\text{slide}} + 0.3 \cdot \mathcal{L}_{\text{inst}} \tag{5.6}$$

A key advantage of CLAM is its interpretability: the attention scores a_k can be visualized as a spatial heatmap overlaid on the original WSI, highlighting the tissue regions that most influenced the slide-level prediction.

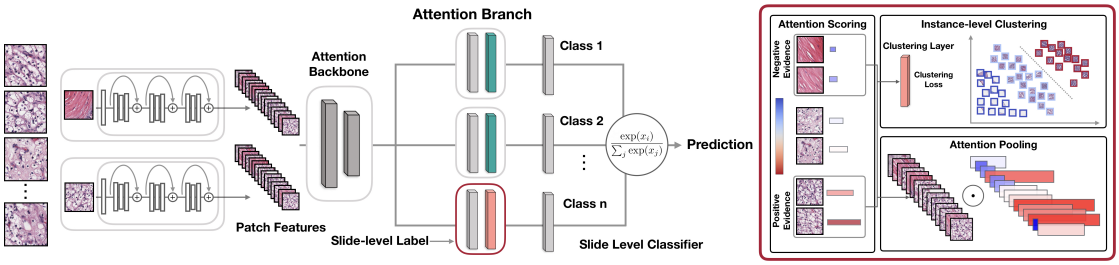

Figure 5.6: Overview of the CLAM architecture adapted from [41]. Patch features are aggregated via a gated attention mechanism that assigns a scalar weight a_k to each patch. The top-k highest and lowest-attended patches are used to compute the instance-level clustering loss $\mathcal{L}_{\text{inst}}$, which refines the feature space alongside the slide-level cross-entropy loss $\mathcal{L}_{\text{slide}}$.

TransMIL

TransMIL [42] is a transformer-based MIL framework that challenges the assumption, made by attention-based methods such as CLAM, that instances within a bag are independently and identically distributed. In practice, patches extracted from the same WSI exhibit strong morphological and spatial correlations: the tissue surrounding a tumor region influences nearby patches, and the spatial arrangement of cell populations carries diagnostic information that is lost when patches are aggregated independently. TransMIL addresses this by modeling pairwise dependencies among all patches via self-attention, as illustrated in (Figure 5.7). The architecture consists of two stacked transformer layers. Each layer follows the standard Multi-head Self-Attention (MSA) design, in which every patch attends to all other patches in the bag, producing a representation that capture global patterns across the slide. To handle the long sequences typical of WSIs, ranging from hundreds to thousands of patches, the quadratic self-attention is approximated using the Nyströmformer [61], which factorizes the attention matrix via the Nyström method and reduces computational complexity from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$. A *Pyramid Position Encoding Generator (PPEG)* is applied after each transformer layer to inject spatial positional information at multiple scales by convolving the

patch token sequence with depthwise convolutions of increasing kernel size, allowing the model to exploit the two-dimensional layout of patches on the slide. These components are jointly referred to by the authors as the Transformer-based Patch Transformer (TPT) module. A learnable class token `[CLS]` is prepended to the patch sequence and updated through the two transformer layers; its final state is used as the slide-level representation and passed to a linear classifier. The model is trained with a standard cross-entropy loss, without any auxiliary instance-level objective.

The heatmaps produced by TransMIL differ from those of CLAM in how they are derived: rather than coming from an explicit attention pooling module, they are derived from the self-attention weights of the `[CLS]` token in the last transformer layer, averaged across all attention heads. Each weight a_k reflects how strongly the class token attends to patch k when forming the final slide-level representation, and can be visualized as a spatial heatmap over the WSI to identify the regions that most influenced the prediction.

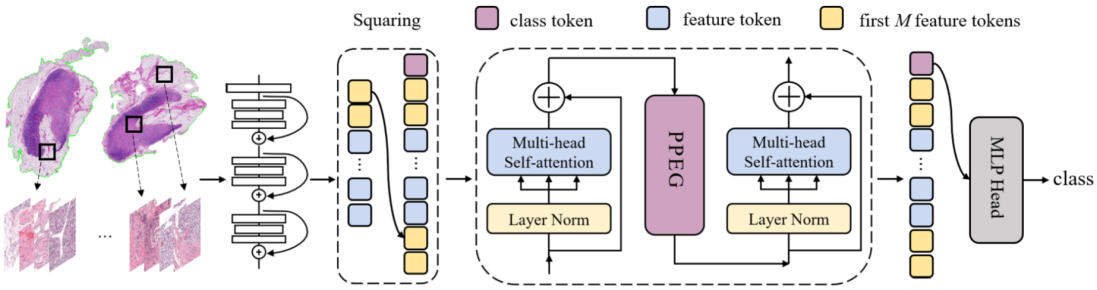

Figure 5.7: The embedded feature vectors are processed by the TPT module through the following steps: (1) Squaring of the sequence to restore the two-dimensional spatial layout; (2) Correlation modeling via Nyströmformer-based self-attention; (3) Conditional position encoding and local information fusion; (4) Deep feature aggregation; (5) Mapping of the aggregated `[CLS]` token representation to the final slide-level prediction.

RRT-MIL

RRT-MIL (Re-embedded Regional Transformer MIL) [43] addresses a limitation shared by all offline feature extraction pipelines: once patch features are extracted by a frozen encoder, they cannot be adapted to the downstream task. RRT-MIL tackles this by adding an online re-embedding step, in which the frozen features are refined by a small transformer trained jointly with the rest of the model.

The re-embedding module consists of two transformer layers, as illustrated in (Figure 5.8). The first applies self-attention within small spatial groups of patches

(*Regional MSA*), so that each patch can exchange information with its neighbors on the slide. The second layer (*Cross-Region MSA*) then lets patches from different groups interact, capturing longer-range context across the slide. Additionally, a lightweight 1D convolution (*Enhanced Positional Encoding Generator (EPEG)*) is applied to inject positional information, helping the model account for where each patch sits on the slide. After re-embedding, the patch features are aggregated by a standard attention pooling module: a scalar weight a_k is assigned to each patch and the slide-level representation is computed as their weighted sum. The model is trained with a standard cross-entropy loss.

The attention weights are used to produce spatial heatmaps over the WSI, in a similar way as CLAM. The authors showed that the re-embedding step can bring the performance of ResNet-50 features up to the level of foundation model encoders, which makes RRT-MIL particularly relevant to the multi-encoder comparison in this work.

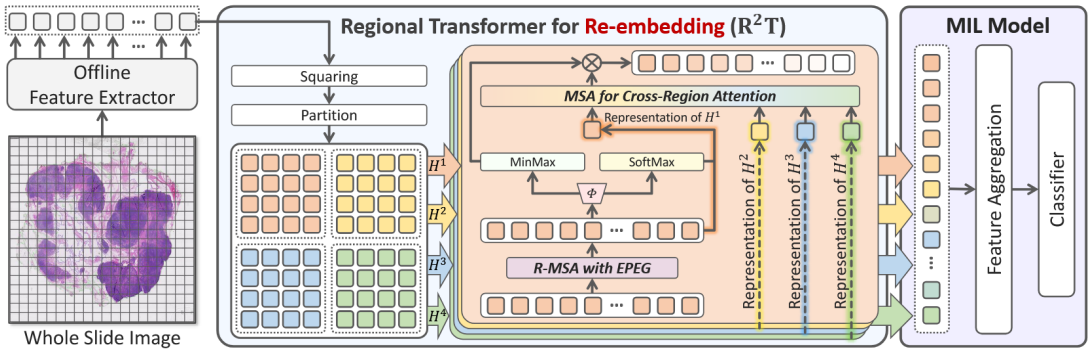

Figure 5.8: Architecture of RRT-MIL, adapted from [43]. Frozen patch features are refined online by two transformer layers: Regional MSA captures local interactions within spatial groups of patches, and Cross-Region MSA aggregates information across groups. A lightweight 1D convolution EPEG injects positional information. The refined features are then aggregated by an attention pooling module for slide-level classification.

Chapter 6

Experiments and Results

This chapter presents the experimental evaluation of both approaches described in the methodology. The supervised approach is evaluated first, followed by the weakly supervised pipeline. For each approach, the experimental setup is described, including the training configuration and evaluation protocol, followed by a discussion of the results.

Performance is reported in terms of AUC, F1-score, and Accuracy, computed at the slide level in both approaches.

All experiments were conducted on the dataset described in Chapter 3.

6.1 Supervised Learning

6.1.1 Experimental Setup

The patch-level classifier was trained on patches extracted from two annotated WSIs: one sarcomatoid mesothelioma case and one pleuritis case, selected from the available annotated slides to maintain a balanced class representation during training. The backbone of the ResNet-18 was kept frozen throughout training, and only the final classification head was fine-tuned for 5 epochs. The training hyperparameters are summarized in Table 6.1.

At inference time, the trained model was applied to all patches extracted from each WSI, and the resulting patch-level malignancy probabilities were aggregated into a slide-level score using the MIL-pool strategy described in Section 4.2. A slide is classified as sarcomatoid mesothelioma if its score exceeds a decision threshold τ, which was varied across the range $[0, 1]$ to identify the value that best separates the two classes.

Hyperparameter	Value
Architecture	ResNet-18
Pretraining	ImageNet
Backbone	Frozen
Epochs	5
Optimizer	SGD
Learning rate	10^{-3}
Momentum	0.9
Loss	Cross-entropy

Table 6.1: Training hyperparameters for the supervised patch-level classifier.

6.1.2 Results

The slide-level classification results are reported in the Table 6.2 below:

	F1 (%)	Accuracy (%)	AUC (%)
Test set	54.05	37.04	28.75

Table 6.2: Slide-level classification performance of the supervised approach on the test set.

The results indicate that the model failed to learn a useful classification boundary, predicting the sarcomatoid class for all slides regardless of the decision threshold τ. Varying τ across its full range did not produce any change in performance, confirming that the issue lies in the score distribution itself rather than in the choice of threshold: the model assigns similarly high probabilities of malignancy to all slides, making the classes indistinguishable at the slide level. An AUC below 0.5 further confirms that the model performs worse than a random classifier, which is consistent with a complete collapse into the majority class.

This outcome is primarily attributable to the severe lack of annotated training data: with only two slides available for training, the classification head does not have enough examples to learn a meaningful decision boundary, and the frozen ImageNet backbone cannot adapt its representations to the histopathological domain. It should be noted that this failure does not invalidate the supervised approach in principle: the pipeline of Naso et al. [2] was designed and validated on a substantially larger annotated dataset, and there is no reason to expect it to generalize from only two training slides. Furthermore, no adjustment to the training configuration, such as modifying the learning rate, the number of epochs, or the aggregation strategy, could be expected to produce a substantial improvement, as

the fundamental bottleneck lies in the scarcity of annotated data itself rather than in any tuneable aspect of the pipeline.

Rather, these results highlight a fundamental practical limitation of patch-level supervised methods: they require numerous pathologist-provided region annotations, which are costly and time-consuming to obtain, and become impractical when only a small number of annotated cases are available. This directly motivates the weakly supervised approach explored in the remainder of this chapter, which relies only on slide-level labels.

6.2 Weakly Supervised Learning

6.2.1 Experimental Setup

The weakly supervised pipeline was evaluated on an internal dataset of 54 slides (20 Sarcomatoid and 34 Pleuritis) using 5-fold cross-validation. The folds were generated through stratified sampling in order to preserve the original class distribution across splits. Due to the inherent class imbalance of the dataset, each validation fold contains a larger number of Pleuritis slides compared to Sarcomatoid ones.

To ensure a fair comparison across models, the same fold assignments were used for all combinations of aggregators (CLAM, RRT-MIL, TransMIL) and feature extractors (ResNet-50, CONCH, UNI v1, UNI v2), so that all experiments were conducted under identical data partitions.

Each model was trained for 20 epochs with a batch size of 1. For evaluation, the checkpoint corresponding to the lowest validation loss was selected. The training hyperparameters adopted for each aggregator are summarized in Table 6.3.

Hyperparameter	**CLAM**	**RRT-MIL**	**TransMIL**
Epochs	20	20	20
Optimiser	Adam	Adam	AdamW
Learning rate	10^{-4}	10^{-5}	10^{-5}
Weight decay	10^{-5}	10^{-5}	10^{-5}
LR scheduler	Cosine	Cosine	Constant
Dropout	0.25	0.25	0.25
Bag weight	0.7	—	—
Inst. loss weight	0.3	—	—
Transformer layers	—	2	2
Attention heads	—	8	8
Transformer dim	—	64	256
Model checkpoint	Minimum validation loss		

Table 6.3: Training hyperparameters for the three weakly supervised aggregators. Dashes indicate that the hyperparameter is not applicable to that architecture.

6.2.2 Shallow Classifiers on Mean-Pooled Features

Before evaluating the full MIL pipeline, a preliminary experiment was conducted to assess the discriminative quality of each feature extractor independently of the aggregation strategy. For each slide, patch-level features were averaged across all patches to obtain a single fixed-dimensional representation. This vector was then used to train a set of shallow classifiers, namely Logistic Regression, Linear SVM, RBF-SVM, and Random Forest.

Model selection was based on balanced accuracy, defined as the mean per-class recall [57]:

$$\text{Balanced Accuracy} = \frac{1}{K} \sum_{k=1}^{K} \frac{TP_k}{TP_k + FN_k} \tag{6.1}$$

where K denotes the number of classes and $\frac{TP_k}{TP_k+FN_k}$ represents the recall of class k. Balanced accuracy is more informative than standard accuracy in the presence of class imbalance, as it assigns equal weight to each class regardless of its prevalence in the dataset. The resulting performance is reported in Table 6.4 and Figures 6.1–6.4.

UNI v1 and UNI v2 achieved the best results, both reaching 92.6% accuracy with zero false positives, whereas ResNet-50 obtained a lower accuracy of 85.2%.

This suggests that pathology foundation models already produce highly discriminative feature representations, where the classes appear well separated in feature space even before applying a learned aggregation mechanism.

Feature Set	**Best Classifier**	**TP**	**TN**	**FP**	**FN**	**ACC**
UNI v2	RBF SVM	16	34	0	4	**92.6**%
UNI v1	Logistic Regression	16	34	0	4	**92.6**%
ResNet-50	Linear SVM	15	32	2	5	85.2%
CONCH	Linear SVM	16	32	2	4	88.9%

Table 6.4: Shallow classifier comparison using mean features. Results refer to the best-performing classifier per feature set, selected by balanced accuracy.

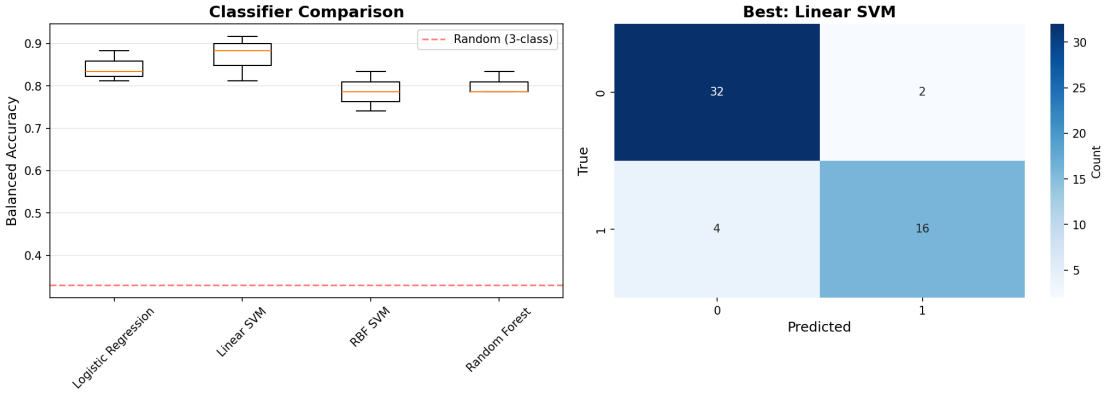

Figure 6.1: Shallow classifier comparison using mean-pooled CONCH features.

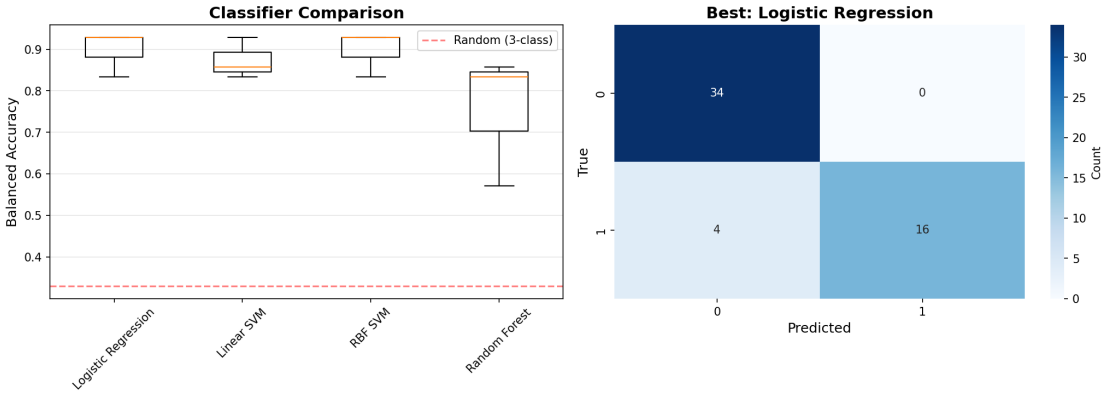

Figure 6.2: Shallow classifier comparison using mean-pooled UNI v1 features.

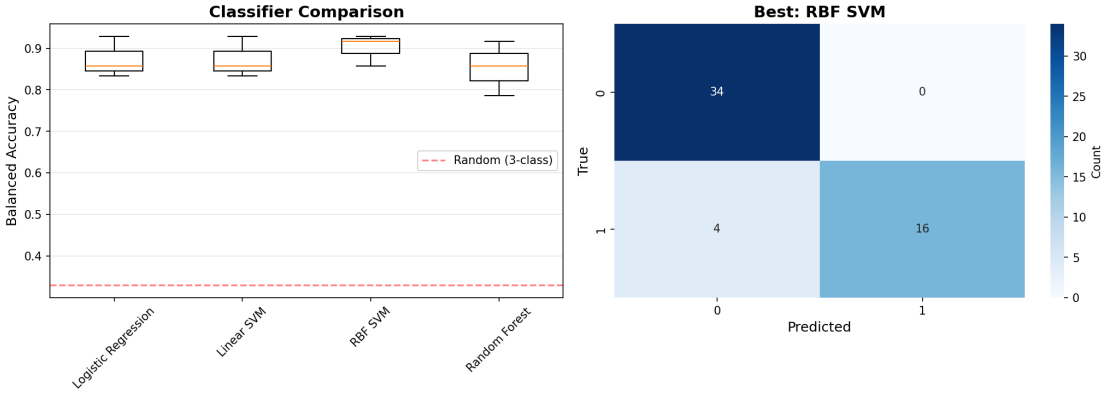

Figure 6.3: Shallow classifier comparison using mean-pooled UNI v2 features.

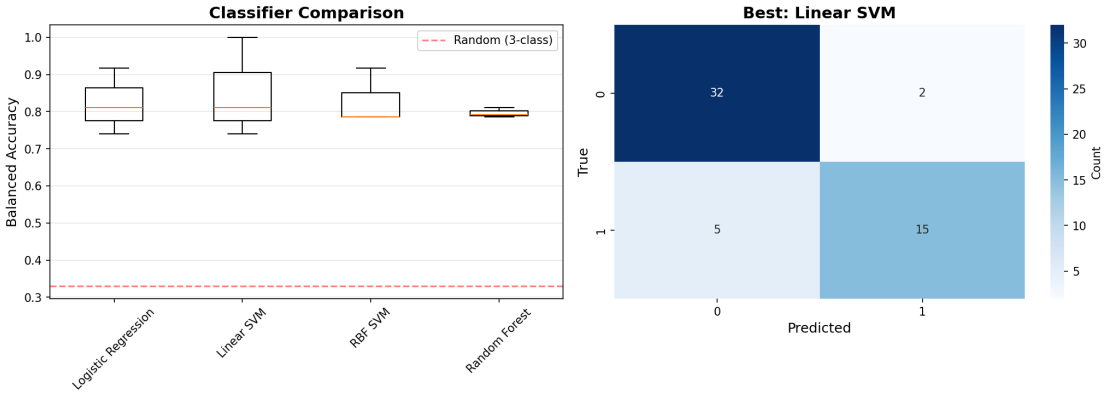

Figure 6.4: Shallow classifier comparison using mean-pooled ResNet-50 features.

6.2.3 Cross-Validation Results

Each aggregator and feature extractor combination was trained and evaluated using the 5-fold cross-validation splits described in Section 6.2.1. At each fold, the model was trained on 43–44 slides and evaluated on the held-out validation set of 10–11 slides. The checkpoint with the minimum validation loss was retained for each fold, and performance metrics were averaged across all five folds. Results are reported as mean ± standard deviation in Table 6.5.

Among the three aggregators, CLAM consistently achieved the highest performance across all feature extractors. Its simpler attention-based pooling mechanism, compared to the transformer-based architectures of TransMIL and RRT-MIL, could have acted as a form of regularization, enabling the model to capture the underlying class distribution without overfitting to the limited training set. In particular, CLAM combined with CONCH yielded the best overall results in terms of both F1 score (0.9714 ± 0.0571) and accuracy (0.9818 ± 0.0364).

Across all aggregators, CONCH performed strongly and consistently, with UNI v1 and UNI v2 closely following. More broadly, all three pathology foundation models outperformed ResNet-50, suggesting the advantage of domain-specific pretraining over generic ImageNet features for this histopathological classification task.

However, these cross-validation results should be interpreted with caution. The internal dataset presents a potential source of bias: all Pleuritis slides were acquired at a single institution, whereas Sarcomatoid cases originate from a different one, with differences in acquisition years and magnification levels. Although several preprocessing steps were applied to mitigate these discrepancies, the models may still partially rely on acquisition-related cues rather than purely pathological features.

The ability of the learned representations to generalize to unseen data will therefore be assessed in Section 6.2.4, where the models are evaluated on an independent external cohort.

Aggregator	Feature Extractor	F1	ACC	AUC
CLAM	ResNet-50	0.8635 ± 0.0127	0.9073 ± 0.0036	0.9357 ± 0.0416
	CONCH	$\mathbf{0.9714 \pm 0.0571}$	$\mathbf{0.9818 \pm 0.0364}$	0.9786 ± 0.0429
	UNI v1	0.9143 ± 0.0700	0.9436 ± 0.0461	0.9929 ± 0.0143
	UNI v2	0.9206 ± 0.0658	0.9436 ± 0.0461	$\underline{1.0000 \pm 0.0000}$
RRT	ResNet-50	0.8571 ± 0.0000	0.9073 ± 0.0036	0.9357 ± 0.0474
	CONCH	$\mathbf{0.8857 \pm 0.0571}$	$\mathbf{0.9255 \pm 0.0374}$	0.9857 ± 0.0286
	UNI v1	0.8381 ± 0.1494	0.9055 ± 0.0856	0.9857 ± 0.0286
	UNI v2	0.7976 ± 0.0775	0.8691 ± 0.0493	$\underline{0.9857 \pm 0.0286}$
TransMIL	ResNet-50	0.7976 ± 0.0775	0.8709 ± 0.0432	0.9714 ± 0.0267
	CONCH	$\mathbf{0.8857 \pm 0.0571}$	$\mathbf{0.9255 \pm 0.0374}$	0.9929 ± 0.0143
	UNI v1	0.8762 ± 0.1227	0.9236 ± 0.0740	0.9774 ± 0.0293
	UNI v2	0.8476 ± 0.1061	0.9055 ± 0.0634	$\underline{1.0000 \pm 0.0000}$

Table 6.5: Validation set performance (5-fold cross-validation). Values are reported as mean ± standard deviation across folds. Best ACC per aggregator is **bold**; best ROC-AUC is underlined.

6.2.4 External Test Set Results

The external test set consisted of five sarcomatoid slides obtained from the Virtual Pathology eLearning repository provided by the University of Leeds [58]. These slides were not used during either training or cross-validation. Since all samples belong to the Sarcomatoid class, performance is reported in terms of true positives (TP) and false negatives (FN) only. Results are summarised in Table 6.6 and Figures 6.5–6.6.

CLAM combined with UNI v1 and UNI v2 correctly classified all external slides (5/5). CLAM with CONCH correctly identified three out of five slides, while CLAM with ResNet-50 detected two out of five. RRT-MIL and TransMIL achieved between two and three correct detections depending on the feature extractor, with UNI v1 and UNI v2 performing best within both aggregators.

Although the small number of external samples prevents drawing strong statistical conclusions, these results provide preliminary evidence that UNI-based features combined with CLAM may offer improved robustness across datasets for sarcomatoid mesothelioma detection.

Aggregator	**Feature Extractor**	**TP**	**FN**	**ACC**
CLAM	ResNet-50	2	3	40%
	CONCH	3	2	60%
	UNI v1	5	0	**100%**
	UNI v2	5	0	**100%**
RRT	ResNet-50	3	2	60%
	CONCH	2	3	40%
	UNI v1	3	2	**60%**
	UNI v2	3	2	**60%**
TransMIL	ResNet-50	2	3	40%
	CONCH	2	3	40%
	UNI v1	3	2	**60%**
	UNI v2	3	2	**60%**

Table 6.6: External test set performance, best Accuracy (ACC) per aggregator is in **bold**.

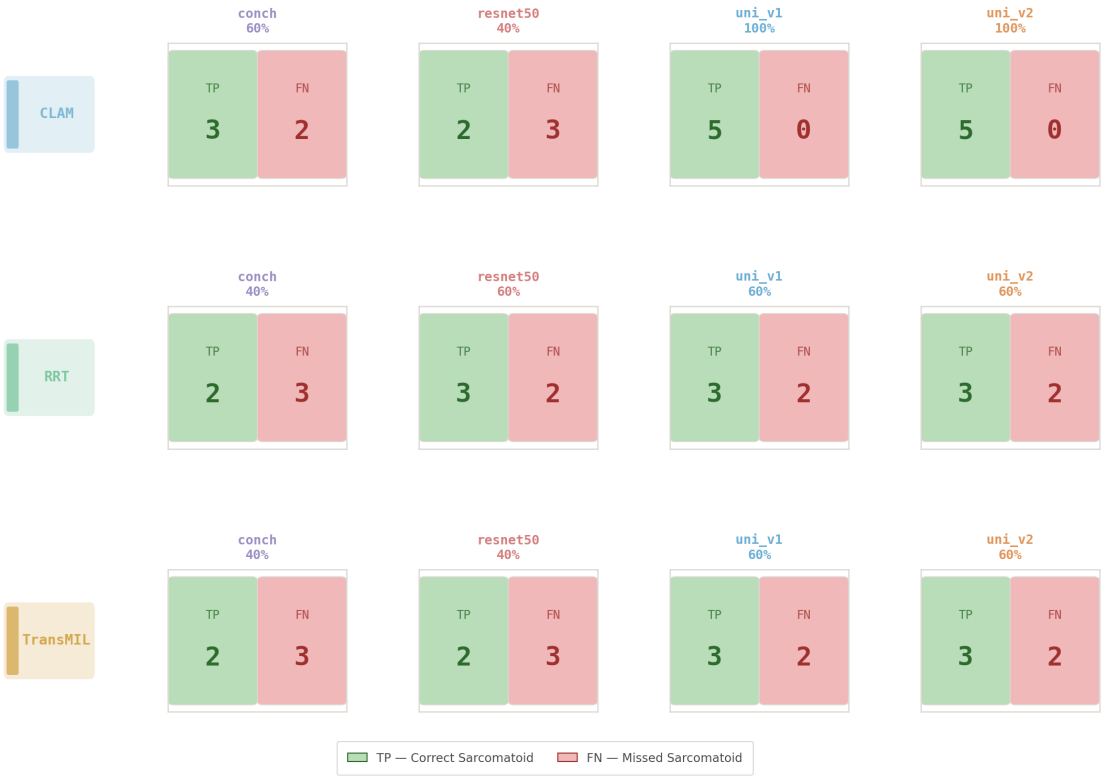

Figure 6.5: External test set confusion matrices for all aggregator–feature extractor combinations.

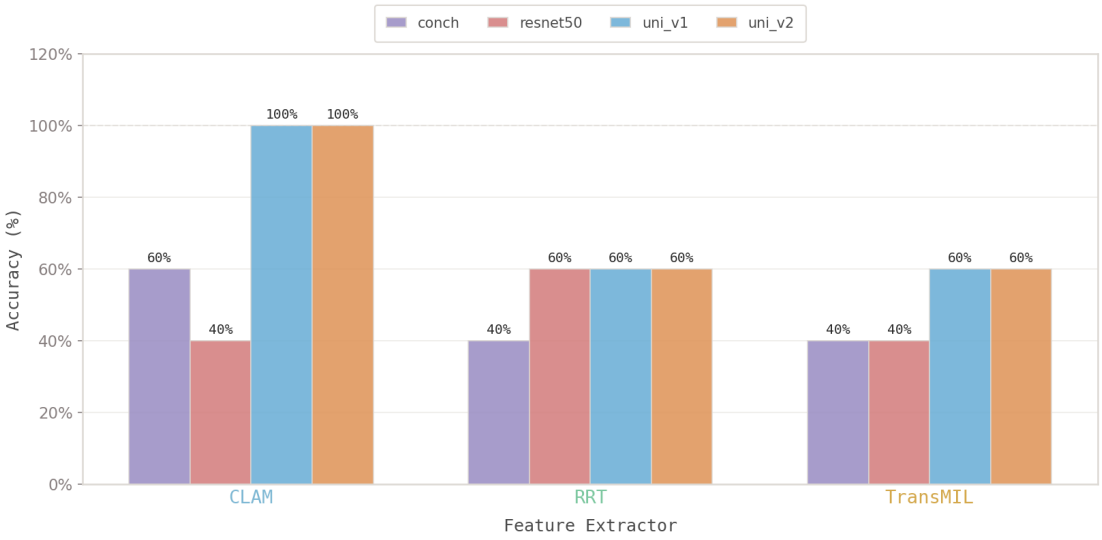

Figure 6.6: External test set accuracy by aggregator and feature extractor.

6.2.5 Attention Heatmap Analysis

Once all aggregators were trained, the best-performing fold for each model was selected to generate slide-level predictions and attention heatmaps across all slides, ensuring that the model used for a given slide had not been exposed to it during training.

Attention scores were percentile-ranked across all patches of each slide and assigned to ten colour-coded bands ranging from the 0–15th to the 95–100th percentile. The top five bands (at or above the 75th percentile) were rendered at full opacity to emphasise the most diagnostically relevant regions, while lower-attention patches were rendered at reduced opacity to preserve spatial context without visual clutter. The heatmaps were additionally exported as ASAPWSI-compatible XML annotations, enabling direct overlay and inspection of high-attention regions within a digital pathology viewer.

The nature of the attention scores differs across aggregators. In CLAM, the scores come from an explicit gated attention pooling module and directly reflect how much each patch contributed to the slide-level prediction. In TransMIL, they are derived from the self-attention weights of the `[CLS]` token in the last transformer layer, averaged across attention heads, and thus reflect global patch interactions rather than a direct contribution score. In RRT-MIL, the scores come from the attention pooling module applied after the re-embedding step, and therefore reflect patch relevance in the refined feature space rather than in the original encoder space. As a result, heatmaps from different aggregators are not directly comparable and should be interpreted in the context of each model's architecture.

A qualitative comparison of the heatmaps against the pathologist annotations was carried out where annotations were available. A full quantitative evaluation was not possible, as complete annotations were only available for a few slides.

Among all combinations tested, CLAM paired with UNI v1 or UNI v2 most reliably highlighted the same tissue regions that the pathologist had marked as tumor or pleuritis, suggesting that UNI features carry enough morphological information to distinguish sarcomatoid mesothelioma from pleuritis. The other combinations showed less consistent overlap with the annotations, though some slides were localized correctly. ResNet-50, in line with its lower quantitative performance, produced the least focused heatmaps, with high-attention regions spread broadly across the tissue rather than concentrated on diagnostically relevant areas.

These observations, combined with the quantitative results on the external test set, point to two main conclusions. First, CLAM generalities better than the other aggregators in this setting, likely because its simpler architecture is less prone to overfitting on a small dataset. Second, UNI-based features consistently produce more informative representations than the other encoders, capturing tissue patterns

that are relevant to the classification even though the model was never trained on pleuritis-specific data. TransMIL and RRT-MIL are more complex models that in principle can capture richer relationships between patches, but this advantage only materialities with enough training data, something that the current dataset cannot provide.

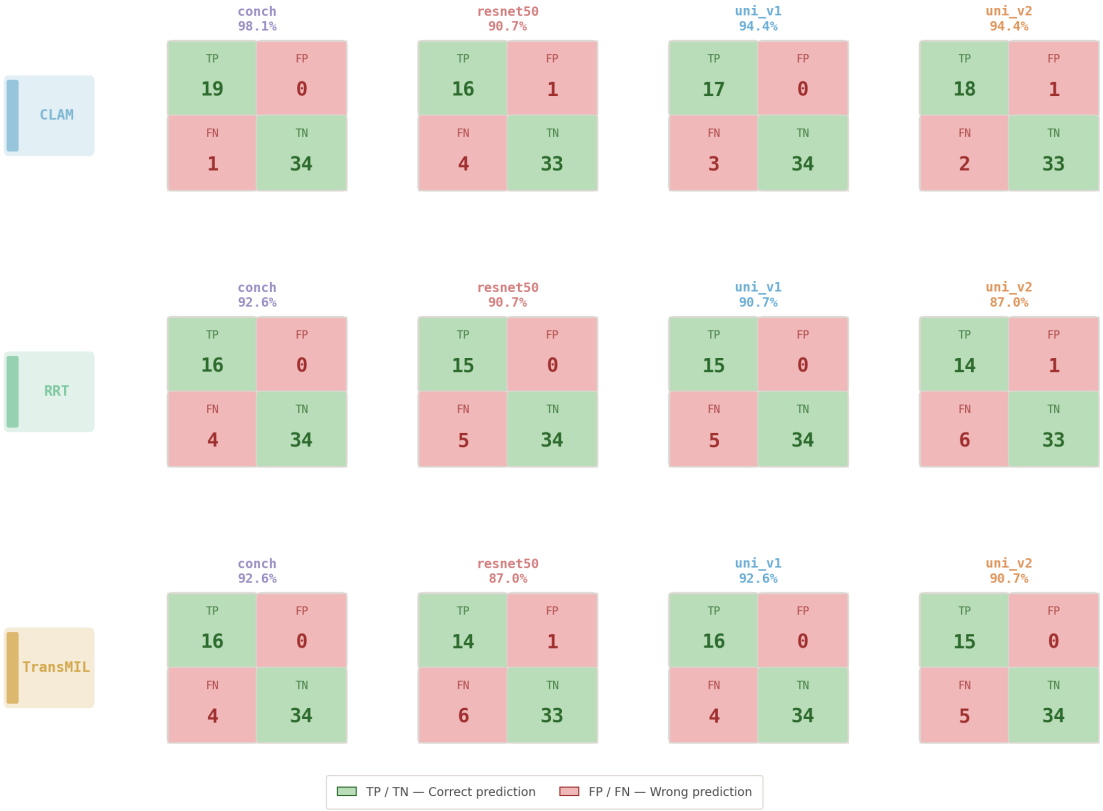

Figure 6.7: Confusion matrices for all aggregator–feature extractor combinations computed when extracting the heatmaps.

6.2.6 Visual Examples

Figure 6.8 shows the pathologist annotations for a representative sarcomatoid mesothelioma slide, which were used as a reference for the qualitative evaluation. Note that this slide was stained with RET rather than H&E; however, since the annotated tumour regions reflect the underlying tissue architecture rather than the staining protocol, the colorimetric differences are not relevant for this comparison. Figures 6.9 and 6.10 show the corresponding attention heatmaps produced by CLAM with UNI v2 and UNI v1 features respectively, illustrating how both models concentrate attention on regions that largely overlap with the annotated tumour areas.

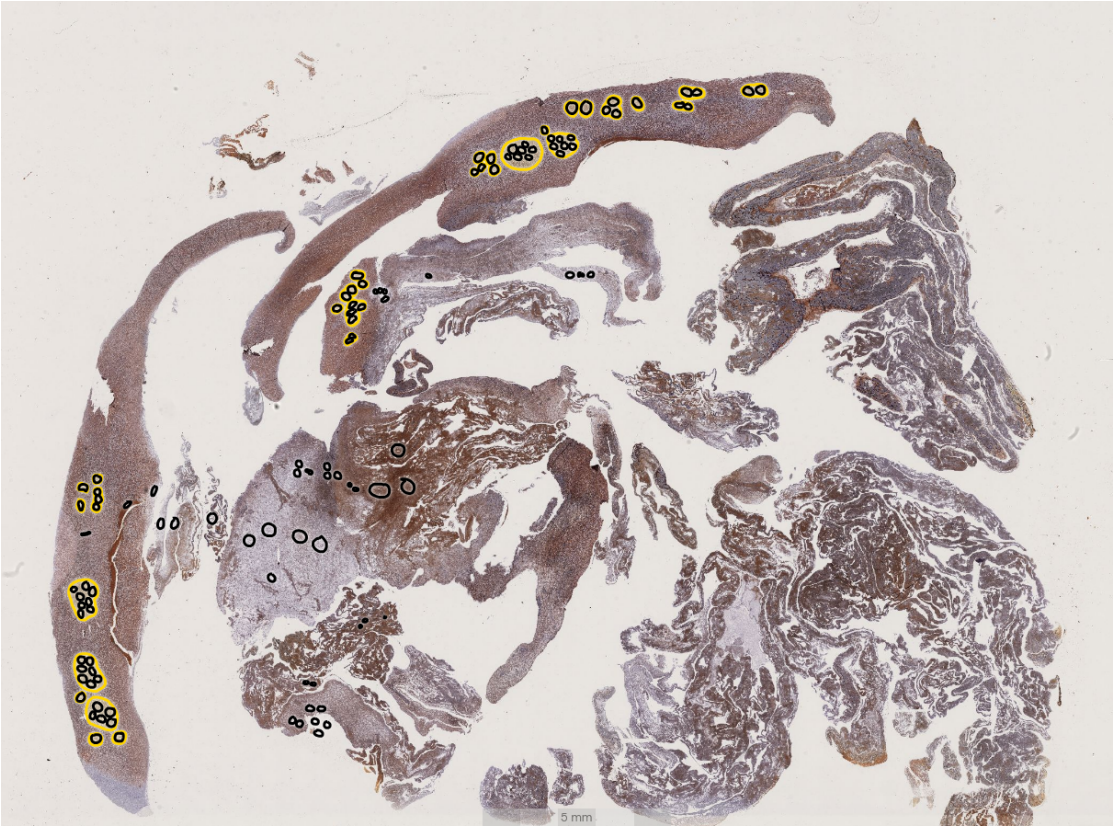

Figure 6.8: Pathologist annotations for a representative sarcomatoid mesothelioma slide, used as reference for the qualitative heatmap evaluation. Only the regions highlighted in yellow correspond to tumor tissue.

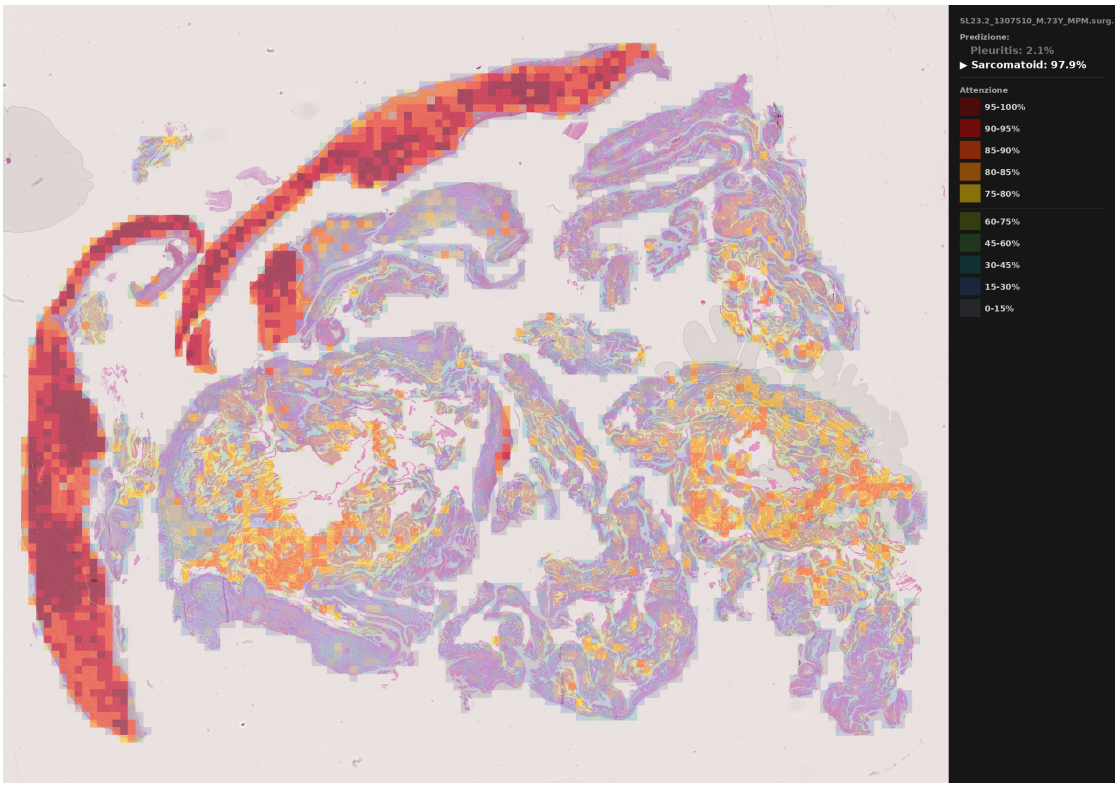

Figure 6.9: Attention heatmap generated by CLAM with UNI v2 features on the same slide. High-attention regions are shown in red.

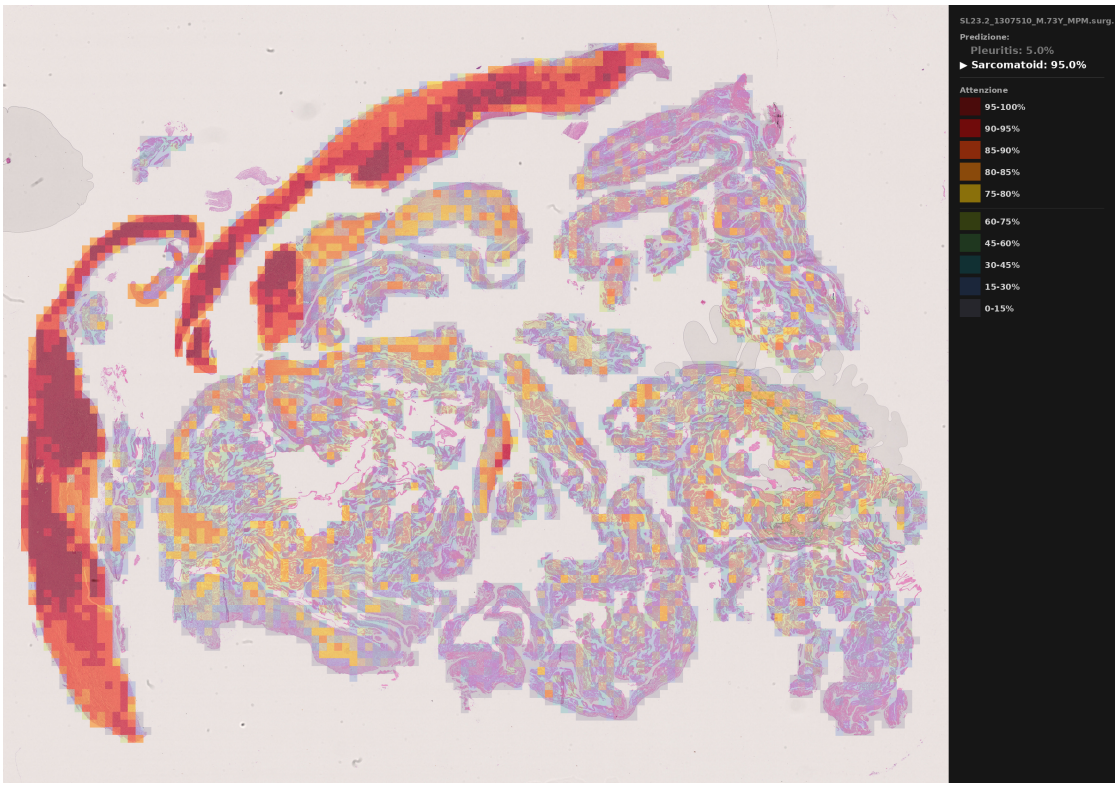

Figure 6.10: Attention heatmap generated by CLAM with UNI v1 features on the same slide. High-attention regions are shown in red.

Chapter 7

Conclusions and Future Works

This thesis investigated the application of artificial intelligence to a challenging diagnostic tasks in computational pathology: distinguishing sarcomatoid mesothelioma from pleuritis using H&E stained WSIs. Two learning paradigms were evaluated on a small, partially annotated, and class-imbalanced dataset that closely reflects the data conditions typically encountered in a real clinical setting.

7.1 Supervised Learning

The supervised approach completely failed to produce a useful classification boundary. The model predicted the sarcomatoid class for all slides regardless of the decision threshold, achieving an AUC of 28.75%, well below random chance. This outcome was not unexpected: the pipeline was originally designed and validated by Naso et al. [2] on a substantially larger annotated dataset, and no adjustment to the training configuration could compensate for having only two annotated training slides.

Beyond this specific failure, the result exposes a more fundamental problem with patch-level supervised approaches. Getting enough precisely annotated patches is hard: pathologists have to manually trace tumor boundaries on high-resolution slides, the process takes a long time, and different experts do not always agree on where exactly the boundaries should be drawn. In a disease as rare as sarcomatoid mesothelioma, all of this becomes even harder and this makes fully supervised models difficult to adopt in practice. WSL, which only requires slide-level labels, sidesteps most of these problems and is therefore a much more practical direction for computational pathology in real clinical settings.

7.2 Weakly Supervised Learning

The weakly supervised approach based on MIL removes the need for precise tissue-level annotations, replacing them with slide-level labels that are far more readily available in clinical practice. This distinction matters: it means that, in principle, any digitized diagnostic slide already carries the information needed to train the model, without requiring additional expert annotation effort.

The results obtained on the internal dataset were strong, and the behavior on the external test set, despite its limited size, was consistent with what was observed during validation. Qualitative inspection of the attention maps reinforced this picture: in many cases, the regions assigned high attention corresponded to areas that a pathologist would consider diagnostically relevant, which lends credibility to the predictions beyond what the numeric metrics alone can convey.

Two factors drove most of the performance differences across configurations: the choice of feature extractor and the aggregation architecture. Among the encoders evaluated, UNI proved to be the most robust. Its large-scale pretraining on hundreds of thousands of WSIs enabled the model to perform well on our challenging dataset, while remaining robust to residual biases that persisted despite preprocessing. CONCH performed similarly well in most settings, while ResNet-50, pretrained on natural images rather than histopathology, lagged behind, confirming that domain-specific pretraining is necessary for this kind of task.

On the aggregation side, CLAM consistently outperformed both TransMIL and RRT-MIL across feature extractors and evaluation conditions. This reflects a recurring pattern in low-data regimes: simpler inductive biases tend to generalize better when training examples are scarce. The attention-based pooling in CLAM appears to act as a natural regularizer, helping the model learn stable decision boundaries without overfitting to the limited training set.

Returning to the central question posed at the beginning of this work:

> *Do weakly supervised learning approaches outperform fully supervised methods in the task of differentiating sarcomatoid mesothelioma from pleuritis using H&E stained WSIs, when only limited and partially annotated data are available?*

The experiments presented here offer a concrete answer. Weakly supervised MIL pipelines are well-suited to this setting and, in our experiments, they outperform fully supervised approaches when only limited and partially annotated data are available. The combination of a pathology foundation model with a lightweight aggregator is sufficient to achieve reliable slide-level predictions even under challenging data conditions. For rare pathologies where large annotated datasets are simply not available, this approach represents a practical and clinically viable starting point.

7.3 Future Works

Several directions could be pursued to build on the findings of this thesis.

Expanding and diversifying the dataset

A key improvement would be the collection of a larger and more diverse dataset, ideally through structured collaboration across multiple hospitals and regions. This would increase both the number of slides and the variability in acquisition conditions, including different scanners, staining protocols, and magnification levels. For rare diseases, federated learning frameworks offer a practical approach to train models on distributed datasets without centralizing sensitive patient data. A larger dataset would also allow more complex aggregation architectures, such as TransMIL and RRT-MIL, to fully leverage their ability to model inter-patch relationships, potentially altering the performance comparison with CLAM observed here.

Few-shot and zero-shot approaches.

An interesting middle ground between supervised and WSL is represented by few-shot and zero-shot methods that exploit the rich representations of vision-language foundation models. CONCH already supports zero-shot image classification through textual queries, and recent frameworks such as FAST [62] have shown that a small number of coarsely annotated examples can guide attention toward diagnostically relevant regions without requiring full patch-level supervision. Exploring such methods in the context of sarcomatoid mesothelioma and pleuritis could be a promising direction, since it would allow the limited annotations available in this dataset to be used more effectively.

Multi-stain and multi-modal integration.

The dataset used in this work contains exclusively H&E stained slides. Including additional staining protocols, such as reticulum staining, or integrating complementary clinical data, for example p16/CDKN2A FISH results which are known to be highly discriminative for sarcomatoid mesothelioma, could provide richer representations and improve robustness. Vision-language models such as CONCH are well suited to this direction, as they can align image features with structured textual descriptions of clinical findings.

Bibliography

[1] N. Kiran et al. «Digital Pathology: Transforming Diagnosis in the Digital Age». In: *Cureus* 15.9 (2023). DOI: 10.7759/cureus.44620 (cit. on pp. 1, 2).

[2] J. R. Naso et al. «Deep-learning based classification distinguishes sarcomatoid malignant mesotheliomas from benign spindle cell mesothelial proliferations». In: *Modern Pathology* 34 (2021), pp. 2028–2035. DOI: 10.1038/s41379-021-00850-6 (cit. on pp. 1, 3, 17, 27, 29, 45, 60).

[3] K. Basak, K. B. Ozyoruk, and D. Demir. «Whole Slide Images in Artificial Intelligence Applications in Digital Pathology: Challenges and Pitfalls». In: *Turkish Journal of Pathology* 39.2 (2023), pp. 101–108. DOI: 10.5146/tjpath.2023.01601 (cit. on pp. 6, 7).

[4] Minjie Cui and Daniel Y. Zhang. «Artificial intelligence and computational pathology». In: *Laboratory Investigation* 101 (2021), pp. 412–422 (cit. on p. 6).

[5] Hamid R. Tizhoosh and Liron Pantanowitz. «Artificial intelligence and digital pathology: Challenges and opportunities». In: *Journal of Pathology Informatics* 9 (2018), p. 38 (cit. on p. 7).

[6] A. Patel, U. G. J. Balis, J. Cheng, Z. Li, G. Lujan, D. S. McClintock, L. Pantanowitz, and A. Parwani. «Contemporary whole slide imaging devices and their applications within the modern pathology department: A selected hardware review». In: *Journal of Pathology Informatics* 12 (2021), p. 5 (cit. on p. 7).

[7] B. Smith, M. Hermsen, E. Lesser, D. Ravichandar, and W. Kremers. «Developing image analysis pipelines of whole-slide images: Pre- and post-processing». In: *Journal of Clinical and Translational Science* 5 (2020), e38 (cit. on p. 7).

[8] Nehal Atallah, Michael Toss, Clare Verrill, Manuel Salto-Tellez, David Snead, and Emad Rakha. «Potential quality pitfalls of digitalized whole slide image of breast pathology in routine practice». In: *Modern Pathology* 35 (2021). DOI: 10.1038/s41379-021-01000-8 (cit. on p. 8).

[9] Seyed Mohammad Mehdi Hosseini. «Uncertainty-aware Renal Cell Carcinoma Subtype Classification». Rel. S. Di Cataldo, F. Ponzio. MA thesis. Politecnico di Torino, Corso di laurea magistrale in ICT for Smart Societies, 2024. URL: https://webthesis.biblio.polito.it/33161/ (cit. on pp. 8, 10).

[10] A. H. Fischer, K. A. Jacobson, J. Rose, and R. Zeller. «Hematoxylin and eosin staining of tissue and cell sections». In: *CSH Protocols* (2008), pdb.prot4986. DOI: 10.1101/pdb.prot4986 (cit. on p. 9).

[11] Mark Zarella, Douglas Bowman, Famke Aeffner, Navid Farahani, Albert Xthona, Syeda Absar, Anil Parwani, Marilyn Bui, and Douglas Hartman. «A Practical Guide to Whole Slide Imaging: A White Paper From the Digital Pathology Association». In: *Archives of Pathology & Laboratory Medicine* 143 (2018). DOI: 10.5858/arpa.2018-0343-RA (cit. on p. 10).

[12] N. Mahabadi, A. Goizueta, and B. Bordoni. «Anatomy, Thorax, Lung Pleura and Mediastinum». In: *StatPearls* (2024) (cit. on p. 11).

[13] E. Agostoni. «Mechanics of the pleural space». In: *Physiological Reviews* 52 (1972), pp. 57–128 (cit. on p. 11).

[14] N. S. Wang. «Anatomy of the pleura». In: *Clinics in Chest Medicine* 19 (1998), pp. 229–240 (cit. on p. 11).

[15] M. Hunter, J. Goldin, and H. Regunath. «Pleurisy». In: *StatPearls* (2024) (cit. on pp. 11, 13, 14).

[16] K. Butler and S. Swencki. «Chest pain: A clinical assessment». In: *Radiologic Clinics of North America* 44.2 (2006), pp. 165–179 (cit. on pp. 11, 12).

[17] F. R. McGuire, T. Gourdin, J. L. Finley, and G. Downie. «Xanthomatous pleuritis mimicking mesothelioma». In: *Respiration* 77.2 (2009), pp. 215–218 (cit. on pp. 13, 14).

[18] K. Hara et al. «Epidemiologic evaluation of pleurisy diagnosed by surgical pleural biopsy using data from a nationwide administrative database». In: *Thoracic Cancer* 13.8 (2022), pp. 1136–1142 (cit. on p. 13).

[19] F. J. Brims, H. Davies, and Y. C. Lee. «Respiratory Chest Pain: Diagnosis and Treatment». In: *Medical Clinics of North America* 94.2 (2010), pp. 217–232 (cit. on p. 13).

[20] M. Rolo, B. González Blanco, C. A. Reyes, N. Rosillo, and P. López Roa. «Epidemiology and factors associated with extrapulmonary tuberculosis in a low prevalence area». In: *Journal of Clinical Tuberculosis and Other Mycobacterial Diseases* 32 (2023), p. 100377 (cit. on p. 13).

[21] B. Clopton, W. Long, M. Santos, A. Asarian, R. Genato, and P. Xiao. «Sarcomatoid mesothelioma: unusual findings and literature review». In: *Journal of Surgical Case Reports* 2022.11 (2022), rjac512 (cit. on pp. 13, 14).

[22] Patrícia Leitão, André Araújo, Bruno Araújo, and José Gonçalves. «Pleural sarcomatoid mesothelioma: a rare type of malignant mesothelioma». In: *Acta Radiológica Portuguesa* 29.2 (2017). DOI: 10.25748/arp.10545 (cit. on pp. 13, 14).

[23] L. T. Nickell, J. P. Lichtenberger III, L. Khorashadi, G. F. Abbott, and B. W. Carter. «Multimodality imaging for characterization, classification, and staging of malignant pleural mesothelioma». In: *RadioGraphics* 34.6 (2014), pp. 1692–1706 (cit. on p. 13).

[24] S. Klebe, N. A. Brownlee, A. Mahar, J. L. Burchette, T. A. Sporn, R. T. Vollmer, and V. L. Roggli. «Sarcomatoid mesothelioma: a clinical–pathologic correlation of 326 cases». In: *Modern Pathology* 23 (2010), pp. 470–479 (cit. on p. 13).

[25] D. Wu et al. «Diagnostic usefulness of p16/CDKN2A FISH in distinguishing between sarcomatoid mesothelioma and fibrous pleuritis». In: *American Journal of Clinical Pathology* 139.1 (2013), pp. 39–46 (cit. on p. 14).

[26] T. Hida et al. «Sarcomatoid mesothelioma with bland histologic features: a potential pitfall in diagnosis». In: *Modern Pathology* 25.5 (2012), pp. 671–680 (cit. on p. 14).

[27] M. Kang et al. «Machine learning classification of pleural lesions using DNA methylation profiles». In: *Nature Communications* 13 (2022), p. 5678 (cit. on p. 14).

[28] Kitajima et al. «Deep learning for malignant pleural mesothelioma diagnosis». In: *Oncotarget* 12 (2021), pp. 1187–1196 (cit. on p. 14).

[29] Laurent Younes. *Introduction to Machine Learning*. 2025. arXiv: 2409.02668 [stat.ML]. URL: https://arxiv.org/abs/2409.02668 (cit. on p. 15).

[30] Lihi Shiloh-Perl and Raja Giryes. *Introduction to deep learning*. 2020. arXiv: 2003.03253 [cs.LG]. URL: https://arxiv.org/abs/2003.03253 (cit. on p. 15).

[31] Keiron O'Shea and Ryan Nash. *An Introduction to Convolutional Neural Networks*. 2015. arXiv: 1511.08458 [cs.NE]. URL: https://arxiv.org/abs/1511.08458 (cit. on p. 15).

[32] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. «ImageNet: A large-scale hierarchical image database». In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848 (cit. on p. 16).

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV]. URL: https://arxiv.org/abs/1512.03385 (cit. on p. 16).

[34] Anant Madabhushi and George Lee. «Image analysis and machine learning in digital pathology: Challenges and opportunities». In: *Medical Image Analysis* 33 (2016), pp. 170–175. DOI: 10.1016/j.media.2016.06.037 (cit. on p. 16).

[35] M. L. Mendelsohn, W. A. Kolman, B. Perry, and J. M. Prewitt. «Morphological analysis of cells and chromosomes by digital computer». In: *Methods of Information in Medicine* 4 (1965), pp. 163–167 (cit. on p. 16).

[36] M. L. Mendelsohn, W. A. Kolman, B. Perry, and J. M. Prewitt. «Computer analysis of cell images». In: *Postgraduate Medicine* 38 (1965), pp. 567–573 (cit. on p. 16).

[37] Jeroen van der Laak, Geert Litjens, and Francesco Ciompi. «Deep learning in histopathology: the path to the clinic». In: *Nature Medicine* 27.5 (2021), pp. 775–784. DOI: 10.1038/s41591-021-01343-4 (cit. on p. 16).

[38] Metin N. Gurcan, Laura Boucheron, Ali Can, Anant Madabhushi, Nasir Rajpoot, and Bulent Yener. «Histopathological Image Analysis: A Review». In: *IEEE Reviews in Biomedical Engineering* 2 (2009), pp. 147–171. DOI: 10.1109/RBME.2009.2034865 (cit. on p. 16).

[39] Michaela Unger and Jakob Nikolas Kather. «Deep learning in cancer genomics and histopathology». In: *Genome Medicine* 16 (2024), p. 44. DOI: 10.1186/s13073-024-01315-6 (cit. on p. 16).

[40] Muhammad Waqas, Syed Umaid Ahmed, Muhammad Atif Tahir, Jia Wu, and Rizwan Qureshi. «Exploring Multiple Instance Learning (MIL): A brief survey». In: *Expert Systems with Applications* 250 (2024), p. 123893. DOI: 10.1016/j.eswa.2024.123893 (cit. on p. 18).

[41] Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. «Data-efficient and weakly supervised computational pathology on whole-slide images». In: *Nature Biomedical Engineering* 5 (2021), pp. 555–570. DOI: 10.1038/s41551-020-00682-w (cit. on pp. 18, 30, 36, 40, 41).

[42] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. *TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification*. 2021. arXiv: 2106.00908 [cs.CV]. URL: https://arxiv.org/abs/2106.00908 (cit. on pp. 18, 41).

[43] Wenhao Tang, Fengtao Zhou, Sheng Huang, Xiang Zhu, Yi Zhang, and Bo Liu. *Feature Re-Embedding: Towards Foundation Model-Level Performance in Computational Pathology*. 2024. arXiv: 2402.17228 [cs.CV]. URL: https://arxiv.org/abs/2402.17228 (cit. on pp. 18, 42, 43).

[44] Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M. Hospedales. «Self-Supervised Representation Learning: Introduction, advances, and challenges». In: *IEEE Signal Processing Magazine* 39.3 (2022), pp. 42–62. DOI: 10.1109/msp.2021.3134634 (cit. on p. 19).

[45] Richard J. Chen et al. «Towards a General-Purpose Foundation Model for Computational Pathology». In: *Nature Medicine* (2024) (cit. on pp. 20, 36, 38).

[46] Ming Y. Lu et al. *Towards a Visual-Language Foundation Model for Computational Pathology*. 2023. arXiv: 2307.12914 [cs.CV]. URL: https://arxiv.org/abs/2307.12914 (cit. on pp. 20, 37, 39).

[47] Maxime Oquab et al. «DINOv2: Learning Robust Visual Features without Supervision». In: *Transactions on Machine Learning Research* (2024). URL: https://openreview.net/forum?id=a68SUt6zFt (cit. on pp. 20, 36).

[48] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. «Transformer-based unsupervised contrastive learning for histopathological image classification». In: *Medical Image Analysis* 81 (2022), p. 102559. DOI: 10.1016/j.media.2022.102559 (cit. on p. 20).

[49] Shekoofeh Azizi et al. *Robust and Efficient Medical Imaging with Self-Supervision*. 2022. arXiv: 2205.09723 [cs.CV]. URL: https://arxiv.org/abs/2205.09723 (cit. on p. 20).

[50] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. *CoCa: Contrastive Captioners are Image-Text Foundation Models*. 2022. arXiv: 2205.01917 [cs.CV]. URL: https://arxiv.org/abs/2205.01917 (cit. on pp. 20, 37).

[51] Yihao Liu and Minghua Wu. «Deep learning in precision medicine and focus on glioma». In: *Bioengineering & Translational Medicine* 8 (2023). DOI: 10.1002/btm2.10553 (cit. on p. 21).

[52] Peter Bankhead et al. «QuPath: Open source software for digital pathology image analysis». In: *Scientific Reports* 7.1 (2017), p. 16878. DOI: 10.1038/s41598-017-17204-5 (cit. on p. 21).

[53] R. Escobar Díaz Guerrero, L. Carvalho, T. Bocklitz, J. Popp, and J. L. Oliveira. «Software tools and platforms in Digital Pathology: A review for clinicians and computer scientists». In: *Journal of Pathology Informatics* 13 (2022), p. 100103. DOI: 10.1016/j.jpi.2022.100103 (cit. on p. 21).

[54] Adam Goode, Benjamin Gilbert, Jan Harkes, Drazen Jukic, and Mahadev Satyanarayanan. «OpenSlide: A vendor-neutral software foundation for digital pathology». In: *Journal of Pathology Informatics* 4 (2013), p. 27. DOI: 10.4103/2153-3539.119005 (cit. on p. 22).

[55] Adam Paszke et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. 2019. arXiv: 1912.01703 [cs.LG]. URL: https://arxiv.org/abs/1912.01703 (cit. on p. 23).

[56] Charles R. Harris et al. «Array programming with NumPy». In: *Nature* 585.7825 (2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2 (cit. on p. 23).

[57] Fabian Pedregosa et al. *Scikit-learn: Machine Learning in Python*. 2018. arXiv: 1201.0490 [cs.LG]. URL: https://arxiv.org/abs/1201.0490 (cit. on pp. 23, 48).

[58] University of Leeds. *Virtual Pathology eLearning*. 2025. URL: https://www.virtualpathology.leeds.ac.uk/ (visited on 03/09/2026) (cit. on pp. 24, 25, 53).

[59] Nobuyuki Otsu. «A threshold selection method from gray-level histograms». In: *IEEE Transactions on Systems, Man, and Cybernetics* 9.1 (1979), pp. 62–66. DOI: 10.1109/TSMC.1979.4310076 (cit. on pp. 27, 32).

[60] K. He, X. Zhang, S. Ren, and J. Sun. «Deep Residual Learning for Image Recognition». In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90 (cit. on pp. 29, 36).

[61] Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. *Nyströmformer: A Nyström-Based Algorithm for Approximating Self-Attention*. 2021. arXiv: 2102.03902 [cs.CL]. URL: https://arxiv.org/abs/2102.03902 (cit. on p. 41).

[62] Kexue Fu, Xiaoyuan Luo, Linhao Qu, Shuo Wang, Ying Xiong, Ilias Maglogiannis, Longxiang Gao, and Manning Wang. «FAST: A Dual-Tier Few-Shot Learning Paradigm for Whole Slide Image Classification». In: *Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS 2024)*. NIPS '24. Vancouver, BC, Canada: Curran Associates Inc., 2024, pp. 3337–3360. ISBN: 979-8-33131-438-5 (cit. on p. 62).