Automatic Classification of Lung Cancers from Histopathology Images

DPhil Thesis

George Batchkala

supervised by

Prof. Jens Rittscher,

Prof. Fergus Gleeson

Big Data Institute, University of Oxford

Hilary 2025

Acknowledgements

Professor Fergus Gleeson has funded me through his A2 research funds throughout my DPhil as part of the DART Lung Health Programme (Innovate UK grant 40255). This has been made possible thanks to the EPSRC Center for Doctoral Training in Health Data Science (EP/S02428X/1), which has awarded me a full Health Data Science Studentship. The computational aspects of this research were funded from the NIHR Oxford BRC with additional support from the Wellcome Trust Core Award Grant Number 203141/Z/16/Z. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Approval for the DART lung health project was granted by the Clinical Trials and Research Governance Committee of the University of Oxford (Number: PID15885-A003-SP001, Date 24/02/2022). DART lung health project has received the following approvals. REC: 21/WM/0278 - 22/12/2021. CAG: 22/CAG/0010 - 22/12/2022. IRAS: 301420 - 24/2/2022.

The results presented in this thesis are in part based upon data generated by the TCGA Research Network `https://www.cancer.gov/tcga`. Some of the data used in this thesis were generated by the National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC) [10]. Ethical approval was not required for TCGA and CPTAC data, as confirmed by the license attached with the open-access data.

I would now like to thank the people who had a positive impact on my DPhil journey in no particular order:

- Anne Powell and Fergus Gleeson, who relentlessly led the DART project forward despite the continuous hindrance from both nature (Covid-19) and people.
- Mark McCole and Cecilia Brambilla, the DART lung health project pathologists, for explaining their workflow to me while we developed and evolved the annotation protocol and for finding the time to annotate my data.

- Nasullah Khalid Alham for managing the data for the AIDA annotation platform.
- Stefano Malacrino for creating and open-sourcing the WSI reader that I used for my first project.
- Korsuk Sirinukunwattana for explaining the workings of whole slide image processing at the very start of my DPhil.
- Bin Li, Mengran Fan, and Yang Hu for first advising, then discussing, and finally collaborating on research ideas.
- Alexander Sauer, Ruby Wood, Louis-Oscar Morel, and Emily Thomas for being good peers throughout my DPhil journey.
- Paul Tourniaire for agreeing to share his MS-CLAM code repository after receiving my email so I do not need to implement it from scratch.
- Konstantinos Kamnitsas and Vicente Grau, my Transfer and Confirmation of Status examiners, for warning me about the dangers of waiting for data and encouraging me to re-evaluate my research plans continuously.
- Jens Rittscher, my group lead and advisor, for being direct and honest but kind and understanding while giving feedback and providing advice during the 4 years.

Last but not least, I would like to thank my family and friends, who made my life enjoyable while I worked on my doctoral project.

Statement of Originality

I declare that this thesis is entirely my own work and, except where stated, describes my own research.

George Batchkala,

Linacre College

Abstract

Lung cancer accounts for more deaths than any other type of cancer. Currently, most lung cancers are diagnosed in symptomatic patients using CT scans and CT-guided biopsies or bronchoscopies. The latter two involve surgical excision of a small piece of tissue. To make a diagnosis, a pathologist examines the tissue under a microscope at different magnifications, noting cytological features and architectural patterns. These observations are aggregated into lung cancer subtypes, which may exhibit multiple characteristic patterns.

My research contributed to the broader DART Lung Health programme, which was based on the Targeted Lung Health Check programme conducted by NHS England. DART's main goals were to generate large datasets to enhance lung cancer diagnosis through quicker, less invasive, and more accurate methods while identifying research opportunities for treatments that could improve survival rates. I worked on automatically classifying lung cancers from histopathology images and creating an annotated histology dataset that would enable connecting histology and CT modalities. My main contributions are:

1. I developed a three-stage protocol for annotating lung cancer histology images from DART. I showed that it is possible to optimise the annotation process by selecting slides or regions with under-represented subtypes or patterns. My work resulted in a multi-centre dataset annotated to the degree unavailable in the public domain.

2. I curated a public lung cancer dataset and proposed using pretext tasks to choose promising patch-level histopathology foundation models for any custom dataset at a fraction of the computational cost of a rigorous benchmarking study. The choice of a good pretext task remains an open avenue of research.

3. I showed that incorporating prior pathology knowledge into model architecture and training pipelines enables models to learn both the dependencies between cancer subtypes and the relative importance of different regions on the whole slide images, improving the lung cancer classification performance as a result.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Contents

This chapter provides the motivation, introduces related work, summarises the main aims and methodologies, and outlines the subsequent chapters of the thesis.

1.1 Lung Cancer Background and Motivation

Lung cancer is accountable for more deaths than any other type of cancer in the world [177]. In the UK, lung cancer is the most lethal (~35,000 deaths p.a.) and third most

common cancer (~48,000 cases p.a.). The number of patients with lung cancer has increased by about 1% over the past decade [6].

Unfortunately, the symptoms of lung cancer often appear when the tumour has spread beyond the lung. Consequently, more than 75% of lung cancers are diagnosed late, and the 5-year survival rate for such late-stage disease is very low at less than 5% [7].

Therefore, early detection and treatment of asymptomatic patients is crucial. Detecting lung cancer early when it is small and seen on a CT scan as a small nodule is now recognised as the best way to do this [61, 173]. These CT scans may be carried out for various reasons: to assess a lung nodule seen on a Chest X-ray (CRX), if a patient has symptoms suggestive of lung cancer, or for other clinical indications such as chest pain; or as part of a lung cancer screening programme. A pooled analysis of 5 lung cancer screening programmes in the UK [26] suggests that lung cancer screening works well in the UK. The average prevalence of lung cancer reported in the pooled study was 2.2%, ranging from 1.8% to 4.4%, i.e., 18 to 44 patients had lung cancer per 1,000 screened patients [26]. The analysis reported the false positive rate of 2% compared to 21% reported by The Cancer Imaging Archive National Lung Screening Trial (TCIA-NLST) [15]. Of patients with a positive result, 53.3% had lung cancer diagnosed. The surgical resection rate for benign cases was 0.07% compared to 24.5% in TCIA-NLST [15].

Once a nodule is detected on CT, it is critical to determine if it is benign, as most of the nodules detected on CT are not due to cancer. About 97% [131] (or 98% [26]) of all nodules seen on CT scans performed for lung cancer screening are not due to cancer and are just benign incidental nodules that cause the patient no harm [131]. Still, they often require multiple scans and sometimes even biopsies and surgery to find out if they are a cancer. The extracted tissue (biopsy or resection) requires a further visual examination under a microscope by a pathologist.

Extracting any tissue from the lungs requires an invasive procedure, which is unpleasant and costly. It also might lead to various complications (see Figure 1.1 for examples of

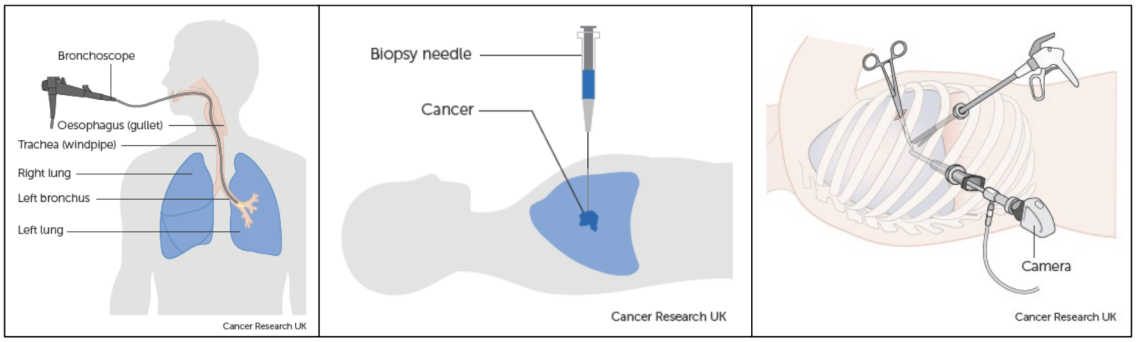

Figure 1.1: Examples of invasive diagnostic procedures for extracting lung tissue. Left: bronchoscopy [57]. Centre: needle biopsy [56]. Right: surgical biopsy [58].

such procedures). Hence, maximising the procedure's value by making faster and more accurate diagnoses is essential for patients and the healthcare system.

The current practice of extracted tissue being examined by pathologists under a microscope can benefit from improvements in determining the general type and the underlying morphological characteristics of lung cancer, since they seriously affect clinical prognosis and potential treatment methods [140].

Some of the common treatment options listed by Cancer Research UK include surgery, radiotherapy, chemotherapy, immunotherapy, targeted drug therapy, and their combinations [5]. Critically, treatment selection is intrinsically linked to histological subtype and molecular markers. For instance, pemetrexed-based chemotherapy [3] is reserved for non-squamous carcinomas [105, 146], while immunotherapy eligibility depends on PD-L1 expression thresholds [22, 138, 200] - both parameters are subject to diagnostic interpretation.

The difficulty in making an accurate diagnosis lies in the inter- and intra-tumour heterogeneity [203]. Furthermore, the trust in the diagnosis from any single pathologist is undermined by the inter- and intra-observer variability among thoracic pathologists [114, 139, 142, 165, 183]. This variability manifests most significantly in these domains:

1. **Subtyping consistency.** Moderate discordance (κ=0.48–0.88) in distinguishing squamous vs. non-squamous Non-Small Cell Lung Cancer (NSCLC) [142, 165],

particularly in poorly differentiated cases [153, 175]. This can lead to inappropriate pemetrexed administration [3].

2. **Biomarker quantification.** PD-L1 scoring near cut-off values shows high inter-observer disagreement (κ=0.32) [22, 138]. PD-L1 scoring errors can deny eligible patients immunotherapy or expose others to ineffective treatment [22, 138]

3. **Staging precision.** Moderate agreement (κ=0.40–0.60) in assessing tumor invasion and nodal involvement [200, 146]. Staging inconsistencies may preclude curative surgery or prompt unnecessary adjuvant therapy [84, 143, 151]

Consequently, diagnostic variability represents a critical juncture in lung cancer care, where histological interpretation directly governs therapeutic efficacy and patient survival outcomes. That makes creating a Computer-Aided Diagnostics (CAD) tool for lung pathology so important.

1.2 AI Applications Background and Motivation

Recently, significant advancements have been made in computer Artificial Intelligence (AI), which have also been utilised in characterising lung nodules detected on CT scans. A promising example of a tool for lung CT images is Optellum's FDA-approved "Virtual Nodule Clinic" [28]. In order to use AI for histology, the slide specimens are digitised.

Digitising pathology slides paves the way for computational pathology tools that improve lung cancer patient pathways. For instance, a computer-aided diagnosis system could triage and prioritise urgent cases lung cancer cases similarly to what has been done for heart transplants [121]. Another tool could highlight diagnostically relevant regions that a pathologist might have overlooked during the examination (leading to better diagnosis). An automated system could request immunohistochemistry-stained slides (resulting in faster turnaround) like done for prostate cancer biopcies [42]. Moreover, predictive models could identify patients likely to respond well to specific therapies [192]. To the

best of my knowledge, no clinically-validated tools for lung histology were available at the start of my doctoral project.

Even though there was no established CAD tool for lung histology, extensive work had been done in that direction over the years. Some researchers [20, 190] explored the ways of determining the predominant pattern of lung adenocarcinoma (the most common type of lung cancer, accounting for about half of all cases [133]). Different predominant patterns determine both the treatment options [116, 157] (the choice of surgery, chemo-, immuno-, or radio-therapy combinations) and the prognosis [30, 88, 196] (best for lepidic, intermediate for acinar and papillary, worst for solid and micropapillary).

Others [55, 117, 124] classified the digitised tissue images into ones with adenocarcinoma and with squamous cell carcinoma - the two most common types of Non-Small Cell Lung Cancer (NSCLC), which in turn accounts for more than 80% of the cases [60, 95]. For non-advanced (early-stage) NSCLC, surgery is the mainstay of treatment for both squamous and non-squamous subtypes. Adjuvant chemotherapy, targeted therapy (in non-squamous with mutations), and radiotherapy are used as needed. The main differences between subtypes are less pronounced in early stages, with targeted therapy being more relevant for non-squamous cases if specific mutations are found [13, 12]. For advanced cases, Squamous NSCLC relies on chemo-immunotherapy combinations with [106] cautious anti-angiogenic avoidance [1], while non-squamous NSCLC leverages mutation-directed targeted therapies [2] and pemetrexed-based regimens [153, 175].

Yang et al. [197] developed a model to classify lung histology images into six different types: 3 most popular lung cancer types (adenocarcinoma, squamous cell carcinoma, small cell lung carcinoma), as well as pulmonary tuberculosis, organizing pneumonia, and normal lung. Compared to NSCLC, Small Cell Lung Cancers have lower survival rates and life expectancy [82].

The drawback of all these methods is that they either look at adenocarcinoma (most prominent lung-cancer type) patterns [20, 190] and do not take other lung-cancer types

into account. Or aim to classify the cancer types directly from the histology images [55, 117, 124, 197] omitting the stage of explicitly finding the morphological features [180] used by the pathologists to make the diagnosis.

My discussions with Dr Mark McCole, a thoracic pathologist at the Oxford University Hospitals NHS Foundation Trust (OUH), suggest that both strategies are unlikely to find support in the pathologist community. Dr McCole believes that pathologists will accept a tool that more closely mimics their workflow: finding diagnostic regions on the pathology slides presenting the WHO-defined features at different magnifications and aggregating them to make the final diagnosis.

1.3 DART Background and Motivation

Currently, most lung cancers are diagnosed in symptomatic patients using CT scans and CT biopsies or bronchoscopies. Recent studies have suggested that screening higher risk individuals provides an opportunity for earlier diagnosis and improved survival [61, 173].

My research was part of the larger project focused on "The Integration and Analysis of **D**ata Using **Ar**tificial Intelligence to Improve Patient Outcomes with **T**horacic Diseases"[1] (**DART**) led by one of my supervisors, Professor Fergus Gleeson. The primary aims of DART were to provide large volume data sets that would enable improvements to the lung cancer pathway by enabling faster, less invasive methods of more accurate diagnosis and suggest areas for research regarding treatment that would improve survival rates.

DART was based on the Targeted Lung Health Check (TLHC) programme[2] conducted by NHS England. This programme invites smokers and ex-smokers aged 55-74 and at higher risk of lung cancer to be screened using a low-dose CT scan at one of multiple sites in England.

During the first stage of the initiative, DART's leadership aimed to get access to 150,000

[1] https://dartlunghealth.co.uk/

[2] https://www.england.nhs.uk/2019/02/lung-trucks/

patients in the TLHC programme to share their medical data, including low-dose CT conducted during screening as part of their standard of care. After their CT scan, if it was positive, the patients selected for follow-up were asked to donate their blood samples and histopathology slides to contribute to the second stage of DART.

One of DART's aims was to use the lessons learnt in developing an AI algorithm for incidentally discovered nodules on CT scans [27] to develop an AI algorithm for nodules detected on lung cancer screening programmes. The AI algorithm had to be specific for lung cancer screening nodules because the patient characteristics are different from those with incidental nodules: incidental nodules occur in a heterogeneous population with lower cancer prevalence [72, 181], while screening targets high-risk cohorts (e.g., heavy smokers) with higher malignancy likelihood [83, 181]. Consequently, nodules may exhibit subtle CT differences due to demographic and risk variations. Additionally, the licensing authorities require the AI algorithm to have been specifically developed and tested on nodules from a lung cancer screening programme.

Another aim was to develop AI for histology to analyse lung biopsies and resection specimens in a similar fashion to the earlier development of AI for CT scans. Notably, when cancer is confirmed at stage I, tumors from both incidental and screening cohorts show no significant differences in size, histology, surgical outcomes, or survival [83]. It is important to make a distinction between healthy, benign, and malignant lung tissue. Healthy lung tissue is characterized by intact alveolar structures. Benign lung lesions, such as hamartomas or granulomas, show organized or disorganized growth patterns (e.g., disorganized cartilage and fat in hamartomas, or inflammatory cells in granulomas) but lack invasion or cellular atypia [140]. In contrast, malignant tissue demonstrates architectural disruption—such as loss of alveolar structure and invasive growth—along with cellular atypia featuring nuclear pleomorphism and high mitotic rates, as well as biochemical changes including glycogen depletion and elevated RNA/proteomic dysregulation [140].

A long-term aim was to combine the CT and histology AI methods and use the new

method to find novel insights from nodules seen on CT scans.

1.4 My Research as Part of the DART Project

This section details my original project aims and their evolution over time.

1.4.1 Original Project Aims

In DART, my supervisor, Professor Jens Rittscher, my colleague Dr Mengran Fan, and I were involved in Work Package 4: "Digital pathology AI and radiomics model development"[3]. Based on the desired improvements to current clinical practice, related literature, and my joint discussions with Professor Fergus Gleeson (DART Principal Investigator, radiologist) and Dr Mark McCole (thoracic pathologist), I have developed specific aims for my project: (1) improve digital pathology Artificial Intelligence (AI) interpretation of nodule and cancer histology from digitised histology slides, (2) develop a radiomics model for detecting the lung nodules and identifying global and fine-grained features on chest CT scans, and (3) explore the connections between histology and radiology identifying new fine-grained radiomics features that could be linked to specific pathological changes. I planned to focus on aim 1 - developing the histology subtyping algorithm, while Dr Mengran Fan would work on aim 2 - the radiology model. In the latter half of my DPhil, we intended to combine our models and work on aim 3 - linking histology and radiology.

One of the primary motivations for the third aim was to identify actionable groups of patients who were scheduled for a biopsy to confirm the suspicion of cancer noted on a CT scan. One group comprises of patients who do not have lung cancer and for whom the biopsy would not be necessary should the CT diagnosis be more precise. The other group consists of patients who could benefit from commencing treatment earlier, without waiting for the biopsy to be scheduled and the subsequent biopsy results.

[3] https://dartlunghealth.co.uk/work-packages/

In order to achieve these aims, our work package was expected to receive pre-operative CT scans from all 150,000 patients, followed by histology slides from 2500 patients. All CT scans were supposed to be accompanied by radiology reports identifying the nodule locations and radiologist observations. All pathology slides were supposed to come with the corresponding pathology reports, leaving only the fine-grained region-level annotations to be done by the Oxford University Hospitals thoracic histopathologist Dr Mark McCole.

1.4.2 Evolution of Project Aims over Time

In July 2020, UKRI announced £11 million in funding for the DART lung health project. I joined the project part-time a year later, in the Summer of 2021 and full-time in September 2021, after completing my Health Data Science CDT training year and joining Professor Rittscher's Quantitative Biomedical Image Analysis research group. DART leadership provided 20 digitised slides from 15 patients at Oxford University Hospitals (OUH) to mitigate the lack of data from DART sites due to COVID-19. With these 20 slides, Dr Mark McCole and I started developing the detailed annotation protocol. In Autumn 2021, we obtained 17 more slides, for a total of 37, and received approval to obtain 200 more slides from OUH.

On Work Package 4, we had an industrial collaborator, Roche. The computational pathology team from Roche had slightly different objectives, and so had their own annotation protocol. We only had Dr McCole's limited annotation time, and because we were unsure whether we could share the detailed annotations due to potential intellectual property rights, he was annotating the slides under both protocols. This meant the detailed annotation of the original 37 slides took approximately 15 months to complete. By Summer 2022, the 37 slides from OUH were fully annotated, but the 200 approved slides had not arrived yet. At the start of 2023, we received 175 slides from OUH.

To speed up the annotation process for this iteration, Dr Cecilia Brambilla, a thoracic

pathologist from Royal Brompton & Harefield Hospitals Guy's and St Thomas' NHS Trust, joined the project. This meant that we needed to split the annotation time of two pathologists with Roche. We reasoned that if Oxford and Roche's legal teams allowed us to share the annotations, it would be better to have annotations for disjoint slide sets. To facilitate this, I developed a tracker sheet system to prevent us from duplicating the annotation efforts. Furthermore, I modified my annotation protocol to make it faster and included only lung cancer subtyping instead of the detailed region-based annotation protocol we employed for the original 37 slides. The subtyping protocol was developed because the pathology reports did not have slide-specific information and referred to the case as a whole. All subtyping annotations were shared with Roche.

By Autumn 2023, we had a total of 212 (37 + 175) slides from OUH with subtype annotations. We also received pre-operative CTs from OUH. Unfortunately, the CTs did not have the nodule locations making them unsuitable for our purposes. With my progress towards developing a histology model being impeded by slow data collection and unusable CT data, Professor Rittscher and I decided to focus purely on the histology part of the DART project (aim 1), leveraging all available public lung cancer pathology datasets for the remaining duration of my DPhil.

At the end of 2023, we started receiving slides from the DART sites, which continued arriving throughout 2024, resulting in 804 slides from 380 patients (including 212 slides from 189 OUH patients) by December 2024. Contrary to our expectations, they came without the pathology reports. This happened because the pathology reports were not adequately digitised and were sent to the data governance team as pictures or scans of printed documents, making the data anonymisation of pathology reports impossible in the time the data governance team had allocated for the DART project. With many more slides coming in and no pathology reports, both the Roche team and I abandoned the detailed annotation and asked the pathologists to focus solely on subtyping as many slides as they could. By January 2025, we had subtyping annotations for 218 DART slides, with most

of the annotations being completed only in January 2025.

We did not receive any CT scans until the Summer of 2024, when 50,000 out of 150,000 chest CT scans arrived. However, due to a data anonymisation problem, the CT scans came without the radiology reports or the tumour locations, making them unusable for our purposes.

1.4.3 Lessons Learnt

To conclude how my research fits into a larger project, I would like to emphasize how projects with ongoing data collection components need solid risk mitigation. Although these projects expose one to the exciting possibilities of developing and modifying your annotation protocols, establishing the datasets not available in the public domain, and conducting research that was never attempted before, they also pose significant risks to making progress and advancements, particularly in the data-driven fields like computational pathology, computer vision, and machine learning.

1.5 Thesis Outline

My thesis has six chapters: an introduction, a literature review, three contribution chapters, and a conclusion.

The first contribution chapter (Chapter 3) introduces the pathology annotation protocol for digitised lung cancer images and its modification over time. This part will be of interest to future researchers working on the annotation of incoming lung cancer data. The protocol contains three consecutive annotation stages, with each subsequent stage taking more time and providing more information about the image: subtyping, region selection, and region annotation. The chapter introduces a method for optimising the region annotation stage given a set of pre-selected regions. The method is described in my first-authored paper, "Active Data Enrichment by Learning What to Annotate in Digital Pathology" [31],

which I presented in September 2022 at the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) Workshop on Cancer Prevention through Early Detection (CaPTion) and in November 2022 at the Digital Pathology (DP) and Artificial Intelligence (AI) showcase organised by Roche at the Royal College of Pathologists. The chapter extends the paper by exploring the possibilities of optimising the other two annotation stages: region and slide selection.

The second contribution chapter (Chapter 4) describes how one can evaluate, select, and use feature extractors pre-trained on pathology images that recently became better known as pathology foundation models. The first part of this chapter comprises my first-authored paper, "Evaluating histopathology foundation models for few-shot tissue clustering: an application to LC25000 augmented dataset cleaning" [34], which received the best paper award at the Data Engineering in Medical Imaging Workshop of the MICCAI 2024 conference. The paper introduces a curated version of a popular tile-level lung and colon cancer dataset. This proposed semi-automatic curation method utilises newly available pathology foundation models and a minimal set-up clustering benchmark for rapid preliminary assessment of new foundation models. The chapter continues by investigating using pretext tasks to narrow the variety of foundation models to a few promising candidates expected to perform best on private clinical datasets.

The last contribution chapter (Chapter 5) focuses on the automatic subtyping of digitised lung tissue into benign, squamous cell carcinoma and adenocarcinoma while also specifying the presence of adenocarcinoma patterns (see Section 2.2.2). The classification was performed in a multi-label setting with partially observed labels on a dataset combined from three public datasets and the DART data. The contributions of this chapter are threefold. First, a method for modelling class dependencies between the cancer types is introduced in my first-authored paper, "Accurate Subtyping of Lung Cancers by Modelling Class Dependencies" [33], which I presented at the International Symposium on Biomedical Imaging (ISBI) 2024 conference and the 2024 British Thoracic Oncology

Group (BTOG) conference [32]. Second, a technique is developed where the annotations from the region selection stage enable the models to pay explicit attention to the diagnostic regions, thereby enhancing prediction precision and interpretability. Finally, a more extensive evaluation of pathology foundation models on an extended dataset is added to conclude this chapter.

1.6 Outputs and Impact

To conclude, I believe lung cancer patients' diagnostics will benefit from my DPhil work. The lessons learnt and recorded here will help to make annotation protocols more efficient, the choice of histopathology feature extractors more rigorous and less computationally expensive, and the modelling approach more closely resembling the pathologist workflow. Furthermore, the advances made to developing models for subtyping lung cancers from histopathology images can be used for the original aim of linking pathology and CT modalities by aligning CT images with cancer types predicted from histopathology images or with slide-level histopathology features.

Incorporating the histopathology-related contributions in a computer-aided diagnosis tool will increase the accuracy and consistency of pathologists' diagnoses from histology images. This, in turn, will be another step on the way to improve patient pathways and reduce patient mortality.

Chapter 2

Literature Review

Contents

Three key areas of knowledge are essential for understanding the methodology presented in the later sections. First, I introduce some terms and ideas from machine learning and modern deep learning. Then, I review concepts from lung pathology. Finally, I dive deeper into the deep-learning approaches that have proved successful in digital pathology tasks.

2.1 Machine Learning and Deep Learning

In my thesis, I primarily focused on classifying lung cancer subtypes and patterns present on digitised lung cancer tissue slides. This task of determining which class best fits a given image is usually referred to as **image classification**.

Image classification models (see Section 2.1.3) take images as input and output the class probabilities for the classes of interest. The task can have multiple flavours, e.g binary, multi-class, and multi-label classification. A binary classification model can determine whether a specific object (cancer subtype) is present or absent in an image. A multi-class classification model can be used to determine which class of objects is the primary focus of the image (cancer type or predominant cancer pattern). A multi-label classification can be used as a combination of the two to answer whether objects of different classes (cancer subtypes and patterns) are present or absent in the image.

Given enough input images with their corresponding class labels, it is possible to develop image classification models that can predict the correct labels for new input images not used in the model development (see Section 2.1.2). Measuring how well the developed model can predict the class of a new input image is done using metrics (see Section 2.1.1).

This section introduces some standard machine learning concepts, deep learning model architectures and blocks, and more recent deep learning concepts and paradigms.

2.1.1 Metrics

Translating what human experts consider good model performance into some mathematical form is a common practice for training machine-learning models. This mathematical form that machine-learning practitioners record and optimise is called a **metric**. In deep learning, differentiable metrics that we want to minimize can also be used as **cost functions**, which are minimized using automatic differentiation and some form of a gradient descent algorithm [155].

For classification problems, the typical metrics used as cost functions are binary and multi-class cross-entropy functions, which I define below. Accuracy can be an intuitive metric, but it is not used as a cost function for deep-learning models because it is non-differentiable.

Set-up: Let the number of samples we are working with be N.

When working with binary classification, let us consider two vectors in $\mathbb{R}^N$:

- $y = (y_1, ..., y_N)^T$ is an $(N \times 1)$ vector of binary labels;
- $\hat{y} = (\hat{y}_1, ..., \hat{y}_N)^T$ is an $(N \times 1)$ vector of model predictions.

Sometimes, we must simultaneously solve multiple (*e.g.* K) binary classification tasks. In this case, let us consider two matrices in $\mathbb{R}^{N \times K}$:

- $Y = (Y_1, ..., Y_N)^T$ is an $(N \times K)$ matrix of binary labels;
- $\hat{Y} = (\hat{Y}_1, ..., \hat{Y}_N)^T$ is an $(N \times K)$ matrix of model predictions.

Each of Y_i and $\hat{Y}_i$ is a $k \times 1$ row-vector in for all $i \in \{1, ..., N\}$.

Binary Cross Entropy (BCE) for sample x_i is calculated as follows.

$$BCE(y_i, \hat{y}_i) = -[y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)] \tag{2.1}$$

For N samples, it can be summarized as a weighted sum of the BCE values of all samples.

Mean-BCE is achieved by making all the weights equal $\frac{1}{N}$, while the Sum-BCE can be obtained by setting all the weights equal to 1.

Multi-class Binary Cross Entropy (Multi-class BCE) is used if for sample x_i we need to predict not 1, but K binary labels. In this case, the metric is calculated as follows for sample x_i and class k.

$$BCE(Y_i^k, \hat{Y}_i^k) = -[Y_i^k log(\hat{Y}_i^k) + (1 - Y_i^k) log(1 - \hat{Y}_i^k)] \tag{2.2}$$

Similarly to BCE, values calculated for each $i \in \{1, ..., N\}$ and $k \in \{1, ..., K\}$ can be summarized as weighted sums for the multi-class case.

Metrics Summary

Ideally, the model's predictions should be as close as possible to the true values of the response. This means that for the multi-class case, we can formulate the following objective for the pair, $Y, \hat{Y}$: Minimize $BCE_{avg}(Y, \hat{Y})$ - average multi-class binary cross entropy.

2.1.2 Dataset Splitting and Model Evaluation

Every time we assess model performance with a chosen metric, *e.g.* with average Global Accuracy (Acc_{avg}), we want some guarantee that the model will be able to generalise, *i.e.* make accurate predictions on previously unseen data. To achieve this, we can imitate the situation of getting new data by splitting the data into train, validation, and test sets. We use them as follows. A **training set** is used to learn from the data, *e.g.* learn the weights of the neural network. A **validation set** is used to perform hyperparameter tuning, *i.e.* to compare models with different hyperparameter settings and choose the best combination. A **test set** is used to imitate unseen data and assess the generalisation performance.

2.1.3 Deep Learning Models and Methods

In this section, I present a brief review of network architectures for image classification, specific architecture blocks, learning paradigms, as well as regularisation and optimisation methods, which have changed the field of computer vision and strongly influenced the work presented in this thesis.

Chapters 3,4, and 5 rely on the methods outlined here in the following ways. First, in all chapters, I used the ResNet [86] (Section 2.1.3) and ViT [68] (Section 2.1.3) models and for extracting features from pathology image patches (Section 2.3.3). These models were pre-trained with Self-Supervised Learning methods (Section 2.1.3), meaning that I used Transfer Learning (Section 2.1.3) when applying them. Second, I used the convolution and attention (Section 2.1.3) blocks in Chapter 5 to communicate the information between different embedding branches of my multi-label classification model. Finally, I applied models pre-trained in a multi-model visual-language fashion (Section 2.1.3) in Chapter 3 for image retrieval and Chapter 5 for image classification.

Network Architectures Overview

Deep learning has come a long way within the field of computer vision since Lecun et al. [115] introduced **LeNet5** for handwritten digit recognition. LeNet5 was the first network that successfully used **average pooling** and **convolution** operations in a practical scenario. It had successive layers of convolution and pooling layers followed by tanh or sigmoid non-linearities. The output of the network was produced by feeding the extracted features into an **MLP block of fully connected layers** - something that is true for all of the successive networks. Since then, the image classification task was the dominant factor for the creation of new **feature extractors** (network parts before the MLP classification head). However, it was not until 2012 when Krizhevsky et al. [111] won the ImageNet competition [62, 156] with **AlexNet**. AlexNet had **successive convolutional layers with kernels of different sizes** (11×11, 5×5, 3×3) and was much larger than LeNet5.

Dropout [90, 164] was used to regularize the network and avoid overfitting. Furthermore, it was trained on a much larger dataset. Hence, its training had to be sped up. The authors achieved this and even utilised a combination of two GPUs and a **ReLU** activation function, which is simpler and faster to differentiate. Another change compared to LeNet5 was the introduction of **max-pooling** instead of average pooling. Since then, the models started getting larger. In 2014, ImageNet challenge researchers from Oxford and Google presented networks powered by new ideas and unprecedented amounts of computational power: **VGG** [163] and **GoogLeNet** (also known as **InceptionV1**) [166]. VGG was much larger than AlexNet and had multiple configurations (from smaller and lighter to larger and heavier). This was the first network to exclusively use 3×3 **convolutions**, suggesting that the successive use of smaller convolutional filters could substitute for a larger convolutional filter at a lower computational cost[1]. The authors of GoogLeNet took the idea of using multiple small convolutional filters even further and created a so-called **Inception block**, which consisted of different sequences of convolutional filters used in parallel (1×1, 3×3, and 5×5 convolutions; 3×3 max pooling). Analogous to the "Network in Network" work [119] from the same year, GoogLeNet also used the idea of implementing 1×1 **convolutions** to reduce the number of features before passing them into the larger convolutional filters. This *bottleneck layer* greatly reduced the computational cost. The depth of the GoogLeNet made it hard to train because the gradients were starting to saturate before propagating from the outputs to the inputs. To counterbalance this, the authors proposed using **multiple auxiliary outputs** at different stages of the network. 2015 was marked by the introduction of **ResNet** [86] - the network that won its authors the 2015 ILSVRC and COCO classification, localisation, and detection tasks. The key improvement He et al. [86] brought to the computer vision community with the ResNet architecture is the use of **residual connections**. They allowed training **very deep networks** (the deepest was ResNet152 with 152 layers), while keeping the computational

[1]This idea did not fully pass the test of time. The extra cost of larger convolution kernels was compensated by efficient implementations and advances in hardware. Furthermore, using a large kernel for the first convolutional layer (e.g., 7×7 [86] or 11×11 [111]) enables capturing low-level features from high-resolution images

cost comparable to the cost of VGG-19 [163]. Similar to VGG, ResNet comes in a variety of configurations, which allows researchers to balance the computational cost and model capacity. ResNet was the first model to beat the human benchmark on ImageNet, achieving a hit@5 error of 3.7% (if one of the top-5 predicted classes is correct, the prediction is deemed correct) and surpassing the human benchmark of 5.1% [107]. I need to add that for the 2015 challenge, Google released two improved versions of GoogLeNet [166] called **InceptionV2** and **InceptionV3** [167]. They showed that a $n \times n$ convolutional filter can be exchanged for a **combination of** $1 \times n$ **n and** $n \times 1$ **filters**, reducing the computational cost per combination and allowing to make the network even deeper. Furthermore, they used **Batch Normalisation** [103] to accelerate training.

Since surpassing the human benchmark, the focus of the ImageNet challenge started shifting towards segmentation, detection, and localization. However, the advancement of classification networks which can be used as generic feature extractors has not stopped. Instead, the efforts split into making more powerful networks which would outdo the performance of their predecessors [52, 94, 93, 168, 193] and creating light-weight networks that would be small enough to run on mobile devices with minimal decline in performance [80, 92, 91, 99, 159, 202, 206]. Having introduced the fundamental architectures and concepts used in this work, I skip the ones cited above and leave their coverage to a review paper or a blog post. The other network worth mentioning is the **EfficientNet** proposed by Tan and Le [170] in 2019. Compared to human-engineered networks, EfficientNet is an example of an architecture found using **Neural Architecture Search** (NAS). Like ResNet, EfficientNet comes in different configurations, from the lightest EfficientNet-b0 to the heavyweight b7 configuration. Like VGG and ResNet before it, EfficientNet is gaining popularity as a baseline network to try on new computer vision tasks.

Transfer Learning

Very few computer-vision labs worldwide have the resources required to train the models described earlier in this section from scratch using either supervised or self-supervised

learning paradigms. Computer-vision practitioners either lack the computational power or the data required for such training. However, a technique exists for applying the knowledge from solving one task with a lot of training data to a new task with less training data called transfer learning. It has been a key to solving many computer vision tasks. The success of transfer learning in computer vision is largely due to the availability of models pre-trained on large publicly-available datasets like ImageNet [62], Pascal-VOC [70], and COCO [120]. The feature extractor parts of these models, excluding the classification head, can be successfully used for downstream classification tasks or as parts of detection, localisation, and segmentation pipelines. These extractor parts can be fine-tuned during downstream training or left frozen (no updates applied during the backward pass).

Capturing Wider Context with Convolutions

I have already mentioned the Inception block in the previous section. Here, I introduce two more blocks used for segmentation and detection tasks, which, therefore, did not make it in the previous section. **Dilated** [198], also known as **Atrous Convolution** [44], [43] is a type of convolutional operation used to capture both *local and global features* on an image. First introduced for semantic segmentation tasks, they can be used to improve the performance of neural networks on fine-grained image data. **Atrous Spatial Pyramid Pooling (ASPP)** [43] is a way to include features at different scales by using Dilated Convolutions with different strides in parallel and concatenating the outputs.

Attention

When speaking about modern deep learning, I can't avoid the topic of attention. The idea of imitating human attention and self-attention mechanisms inside a neural network was popularised by Bahdanau et al. [25] and Vaswani et al. [184]. The ideas and their practical implementations were originally applied to the machine translation task but have, in later years, changed the fields of natural language processing and computer vision. Bahdanau et al. [25] introduced the **attention mechanism** as a practical way to weigh

the importance of the words in a source language for each subsequent translated word in the target language. Vaswani et al. [184] extended the idea to **self-attention mechanism** introducing the **transformer architecture**. This work showed that re-weighting each word in the source language sentence based on its context helps translate the next word in the target language.

The idea of paying different attention to parts of the input depending on this part's features in a wider input context found its way to computer vision and manifested itself in an extension of the original transformer paper [184] into Visual Transformers (**ViT**) [68], where small image patches were considered to be input parts instead of words like in the original transformer work [184].

Self-Supervised Learning

Training deep learning models in a supervised fashion brought significant advancements to the field of computer vision (see Section 2.1.3 for the examples). The supervised learning paradigm, however, has one major flaw - by definition, it requires a dataset with explicit labels. Frequently, these labels need to be provided during the process of manual annotation, which limits the size of the datasets and imposes a significant cost on the dataset creation. The cost is greater for domains which require expert annotations, *e.g.* only qualified doctors can annotate specific datasets in the medical domain.

One of the solutions proposed by researchers in the computer-vision domain is to pre-train networks on larger datasets which do not require manual annotations using "pretext" tasks, *i.e.* tasks that are not of interest on their own, but which force the networks to learn useful representations of the data that can, later on, be used on downstream tasks directly or using transfer learning. Such pre-training falls under a self-supervised or unsupervised learning paradigm since no additional human supervision is required.

Within the domain of self-supervised methods for learning visual representations, **contrastive learning** [85] became popular in 2020 and 2021. Contrastive learning is based

on the idea that inputs $\{x_{1i}, .., x_{1n}\}$ generated by perturbations of some specific input x_1 should have a similar representation to this specific input x_1. $\{x_{1i}, .., x_{1n}\}$ are called positive samples). At the same time, inputs that were not generated but were in the dataset on their own $\{x_2, .., x_m\}$, should be different from the specific input x_1. $\{x_2, .., x_m\}$ are called negative samples. Rather than providing a comprehensive review of contrastive learning methods, I will focus on introducing the approaches I have encountered most frequently in practice: **SimCLR** by Chen et al. [49] and **MoCo** by He et al. [87]. Both works use data augmentation to generate a positive sample from an image. Negative samples are derived from other images; however, in SimCLR, they are drawn from the same batch, whereas in MoCo, they are sourced from a dynamically updated dictionary of samples, which is built as a queue. Each time, a new mini-batch is enqueued, and the oldest mini-batch is dequeued. This decouples the dictionary from the mini-batch size, allowing the dictionary to be larger and more representative of the dataset than any one mini-batch. SimCLR demonstrates that using larger batches and longer training benefits the model, while MoCo utilises momentum to smooth the training process by dynamically updating the dictionary. Both works have been improved since their original publication. **MoCo-v2** [51] uses the ideas from the original SimCLR, an MLP projection head, and more data augmentation to improve compared to both original papers [49, 87]. **SimCLR-v2** [50] shows improvements from using bigger feature extractors than in the original SimCLR.

Following the success of BERT [65], which popularised masked language modelling, the computer vision community adopted **masked image modelling** pretext task to train strong feature extractors: BEiT [29], iBOT [204], DINO [41], and DINOv2 [59, 141]. BEiT was the first work that made the self-supervised pre-training of Vision Transformers (ViTs) outperform supervised pre-training on the ImageNet-1K dataset [156]. All methods utilise knowledge distillation [89] and are primarily used to pre-train Vision Transformer [68] models.

Multi-modal Pretraining

While studies in Section 2.1.3 focused on training models on pretext tasks instead of supervised image classification pre-training, other studies proposed to utilise the dense information provided by the text captions. I briefly describe two studies to showcase the different ideas of vision+language pre-training: VirTex [64] and CLIP (**C**ontrastive **L**anguage–**I**mage **P**re-training) [149]. In VirTex, Desai and Johnson [64] used the COCO [120] dataset of images with captions to learn useful visual features for generating captions. In CLIP, Radford et al. [149] created a dataset of positive and negative pairs of images and captions and used contrastive pre-training to pull together the image and caption embeddings of positive pairs and push apart the image and caption embeddings of negative pairs. While VirTex can be used to generate captions for new images, CLIP pre-training enables zero-shot classification of images and retrieval of images similar to a caption from a given pool of images. To obtain the classification predictions for a given image, one needs to (1) compute the image embedding using CLIP image encoder, (2) compute the embeddings for multiple captions that represent the classes (e.g. "An image of {class C_i}" for each of the classes of interest) using CLIP text encoder, (3) compute the inner products of image and caption embeddings, and (4) apply softmax to calculate probabilities of this image belonging each of the classes. To use CLIP for image retrieval of a specific class, one reverses the situation described above and uses a single caption with multiple images.

2.2 Introduction to Lung Pathology

Potential lung cancers are usually picked up from chest X-ray or CT scan. After that, the lung cancer diagnosis needs either to be confirmed or rejected. If confirmed, different treatments need to be considered based on the type of lung cancer (see Section 1.1). The confirmation and subtyping processes usually require a piece of tissue to be extracted from the potential cancer region. This is usually done through bronchoscopy [57], needle

biopsy [56], or surgical biopsy [58]. After that, the tissue is prepared and presented to a pathologist. The pathologist's conclusions can inform the necessity and choice of (1) the pre-operative treatment e.g. radio-, chemo-, targeted drug, or immuno-therapy and (2) the surgery. If the surgery is performed, the resected tissue is again examined by the pathologist to determine the post-operative treatment (same options as for pre-operative treatment).

The preparation of lung histology slides includes staining with haematoxylin and eosin stain (**H&E**) [11] to highlight tissue and cellular architecture. Hematoxylin stains cell nuclei a deep blue or purple, clearly highlighting nuclear detail. Eosin stains the cytoplasm, extracellular matrix, and connective tissue in varying shades of pink to red. Sometimes, an adjacent section of tissue is also stained with Elastic Verhoeff's Van Gieson stain (**EVG**) [100] to highlight elastic fibres and collagen in tissues: elastic fibres are stained black, collagen fibres are stained red, while other tissue elements (cytoplasm, background) are stained yellow. See Figure 2.1 for examples of differently-stained biopsy and resection images and Section 2.2.1 for details on the preparation process. The pathologist then examines the tissue under a microscope or through a digital slide viewer to determine the presence and quantity of morphological characteristics vital for making the diagnosis. A digitised slide is called a **whole-slide image** or **WSI**.

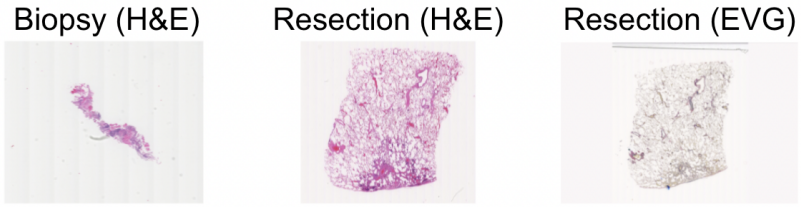

Figure 2.1: Examples of different digitised whole-slide images with different stains.

I now introduce the aspects of lung pathology that are essential for understanding this thesis by walking through the process of tissue preparation and digitisation. Finally, I list the morphological patterns used by pathologists to make a diagnosis.

2.2.1 Tissue Preparation: Sectioning, Staining and Scanning

After a biopsy or tumour removal, the extracted tissue must undergo multiple steps before it is ready for pathologists and researchers to view.

First, the tissue needs to be fixed, which is usually done with 10% neutral buffered formalin. Then, it is transferred to an appropriately-sized tissue cassette before the processing takes place. The processing consists of 3 stages: dehydration (removal of water with alcohol), cleaning (removal of alcohol with an organic solvent), and embedding (infiltrating the tissue with an embedding agent – usually paraffin wax).

After all these steps, the tissue must be thin enough for the light microscopy to work. This is achieved with a *microtome*, a device that utilises a knife/blade or a laser to cut thin sections of tissue. The standard width of a section is 5 μm = 0.005 mm or 5 microns. Finally, most cells are transparent and appear almost colourless when unstained. This is where the staining methods, like H&E or EVG staining help. For more information about the whole process, see [144].

At this stage, pathologists equipped with a light microscope can use the glass slides. However, for the purposes of this project, the slides also needed to be digitised. This was done using specialised scanners, which combine a high-resolution camera with optical magnification. Two standard magnifications used are 20x and 40x with standard resolutions of 0.5 μm per pixel resolution and 0.25 μm per pixel resolution, respectively. The scans presented in this thesis were made with a Hamamatsu digital slide scanner with a 20x lens magnification and a Ventana DP200 scanner with lens magnifications of 20x and 40x. Note, the convention to refer to the 20x and 40x lens magnifications is usually an oversimplification of using a fixed 10x lens (most common fixed lens magnification) in addition to a 20x or 40x variable lens, which means that the actual magnification can be 200x or 400x.

2.2.2 Classification of Lung Tumours

In this section, I present the classification of lung tumours based on the morphological patterns that can be observed on the histology slides. This classification is based on the 2021 WHO Classification of Lung Tumours [140], which was refined by Dr Mark McCole to include all but the very rare lung cancers. Since I aim to use only the imaging-based features, I ignore all immunohistochemistry (IHC) and genetic markers outlined in [140]. Note, unlike H&E and EVG stains that highlight general tissue structures and components such as nuclei, cytoplasm, collagen, and elastic fibres without targeting specific molecules, immunohistochemistry (IHC) stains use antibodies to detect and visualise specific proteins within tissue samples.

The process of diagnosing a tumour from a histology slide follows: The presence, absence, and extent of specific morphological features at different magnifications are noted, and the diagnosis is made when there is enough evidence to classify the specimen into one of the categories based on the features observed. In this section, I take a top-to-bottom approach to presenting all the features typical for different types and subtypes of cancer. In practice, pathologists take a bottom-to-top approach since they do not know the diagnosis in advance.

Benign vs Malignant Tumours. There are several patterns which are usually present in benign tumours. However, these patterns can also be present in malignant tumours. Hence, the tumour will be deemed benign if no malignant patterns are present.

Malignant Tumours. Malignant Tumours can be of three major types: Non-Small Cell Carcinoma (NSCC), Small Cell Carcinoma (SmCC), and Carcinoid Tumour. In turn, these types have different subtypes. Non-Small Cell Carcinoma (NSCC) lung cancer type includes Adenocarcinoma with all its subtypes (AIS, MIA, Invasive A., Mucinous A.), Squamous Cell Carcinoma (SqCC), and Large Cell Neuroendocrine Carcinoma (LCNEC). Small Cell Carcinoma (SmCC) does not have subtypes. Carcinoid Tumours can be typical and atypical. Table 2.2.2 shows abbreviations, short names, and descriptions of

the cancer types and subtypes mentioned here. This list is not exhaustive, but it includes the classification of all but the most rare types and subtypes of lung cancer.

Abbreviation	Full name or/and description
NSCC	Non-Small Cell Carcinoma
A.	Adenocarcinoma
AIS	Adenocarcinoma in-situ: purely lepidic growth, <3cm
MIA	Minimally-invasive Adenocarcinoma: invasive foci <5mm, tumour size <3cm
Invasive A.	Presence of patterns other than lepidic >5mm diameter
Mucinous A.	Cells with abundant intro-cytoplasmic mucin
SqCC	Squamous Cell Carcinoma
LCNEC	Large Cell Neuroendocrine Carcinoma
SmCC	Small Cell Carcinoma

Table 2.1: Lung cancer types: common abbreviations and descriptions.

As mentioned before, different patterns at different image resolutions can assist in differentiating these tumours.

Cytological Features. At high power (high magnification), pathologists focus on *cytonuclear* patterns (cytoplasm and nucleoli). Here I present different subtypes of cancer and the cytological features that help pathologists distinguish between these subtypes.

NSCC	SmCC
Adenocarcinoma pattern: micropapillary	Little Cytoplasm
SqCC: inter-cellular bridges	Stippled/salt & pepper chromatin
LCNEC: mitotic count >10 per 2 mm^2	Indistinct nucleoli
Mucinous A.: intracytoplasmic mucin	Nuclear Moulding
Abundant cytoplasm, prominent nucleoli	

Table 2.2: Cytological features of different lung cancer subtypes.

Stippled/salt & pepper chromatin is also a cytological feature of **carcinoid tumours**.

Architectural Patterns. At low power (low magnification), pathologists focus on *architectural* patterns. Similarly to what I did with high-power patterns, I present different cancer subtypes with their corresponding low-power patterns. I also added descriptions of the adenocarcinoma growth patterns.

LCNEC architectural patterns include islands with central necrosis, peripheral palisading,

Adenocarcinoma	Description
Lepidic	Cytologically malignant cells grow along alveolar walls
Acinar	Malignant glands within a fibroinflamatory stroma
Papillary	Fibrovascular cores covered by cytologically malignant cells
Solid	Sheets or solid islands of non-small cell carcinoma (needs a presence of another adenocarcinoma patterns or immunohistochemistry, e.g. TTF1/Napsin A or histochemical (DPAS) confirmation of diagnosis)
Micropapillary	Small clusters of cells lacking fibrovascular cores within stroma or alviolar spaces (needs confirmation at higher power)

Table 2.3: Architectural adenocarcinoma patterns of NSCC. EVG staining can assist with the distinction between these patterns.

rosettes, and trabeculae. For LCNEC, the final diagnosis requires immunohistochemical confirmation with at least two out of three positive IHC markers (Synaptophysin, Chromogranin A, and CD56) [63]. Additionally, necrosis is one of the typical patterns present in SmCC, LCNEC, and Atypical Carcinoid tumours.

Prognostic Features

Some of the features can assist with predicting patient prognosis. Tumour subtype, size, and grade significantly impact the prognosis. Tumour grade is a measure of how abnormal cancer cells look under a microscope compared to normal cells, indicating how quickly the tumour is likely to grow and spread. For Adenocarcinoma lung cancer, the growth pattern plays an important role: tumours with solid and micropapillary predominant patterns are usually associated with a worse prognosis than tumours with acinar and lepidic patterns. The spread and invasiveness of the tumours are also important. Sometimes, single cells and micropapillary clusters spread within airspaces beyond the contour of the tumour; this is called Spread Through Air Spaces (STAS) and can be seen at high magnification. Pleural and Lymphovascular invasions are the other two criteria affecting the prognosis. EVG staining can assist with determining the presence of the former.

2.3 Advances in Computational Pathology

The space of lung cancers is complex and heterogeneous, as seen in Section 2.2. This is why creating a computer-aided diagnostics (CAD) tool that assists pathologists in making faster, better, and more consistent decisions is essential. Progress has already been made in subtyping the most popular type of lung cancer - adenocarcinoma [20, 190], identifying the general type of lung cancer [55, 117, 124, 197], and predicting mutations [55]. However, to the best of my knowledge, no work explored a wider range of morphological patterns and covered more lung cancer types when I started working on my doctoral project. Focusing on both morphological patterns and different lung cancer subtypes is important because they determine both prognosis and potential treatments (see Section 1.1).

In this section, I present the challenges specific to computational pathology, outline and compare the current methods, and highlight the gaps in research that still need to be filled.

2.3.1 Challenges

Despite the advances that deep learning has brought to computer vision, the applications of deep learning methods to computational pathology have not yet reached the level of adoption comparable to that of computer vision systems in the natural image domain. This can be attributed to the specific challenges that researchers encounter when working with pathology images, which include, but are not limited to, regulatory restrictions on data sharing, the high cost of expert medical annotations, high memory requirements to store gigapixel pathology images, and the need for strict multi-centre clinical validations.

The advances in applied computer vision on natural images can be partially attributed to the availability of model architectures and weights pre-trained on large-scale datasets, such as [54, 62, 70, 120], which enabled transfer learning. The drastic *domain shift* from natural images to pathology images suggests that pathology-specific representations of input images need to be learned, which inevitably requires training data. Towards the

end of my doctoral project, this difficulty has been partially alleviated by the release of pathology-specific foundation models (Section 2.3.4). However, restrictions on data sharing in the medical domain remain an issue for research groups that lack vast amounts of pathology data.

The second factor that can claim part of the credit for the advances in computer vision on natural images is the low cost of annotation, which has allowed the aforementioned datasets to be fully labelled, providing dense supervision for training. The *cost of annotating pathology training data is much higher* compared to natural imaging tasks, as the annotations can only be completed by experts with many years of training and cannot be crowd-sourced.

Finally, *pathology images are much larger than natural images* that were used during the development of the best-known architectures [86, 111, 163, 166, 170] used on natural images (both height and width <1000 pixels). Digitised pathology images have a height and width greater than 10000 or even 100000 pixels. Hence, for the same architectures to be used, the images need to be split into patches before being fed to neural networks for feature extraction. Training these networks on patches directly requires patch-level labels, which are most commonly unavailable due to the prohibitive nature of patch-level annotations. If patch-level labels are unavailable, whole-slide-level labels are used to train the models. This means that patch-level features or predictions need to be aggregated into whole-slide predictions. Authors of several works [39, 124, 176] claimed that their models learnt to recognise clinically relevant features when training on WSIs with WSI-level labels. Validating these claims is not hard and can be done by a pathologist. However, when no explicit pathologist validation is provided, one can only support this claim using post-hoc interpretability methods. When only WSI-level labels are available, the classification of whole-slide images is usually formulated as a multiple-instance learning (MIL) problem.

2.3.2 Multiple-Instance Learning

As mentioned before, the Multiple Instance Learning (MIL) formulation of the task arises when the label is only available for a WSI (bag), which is split into patches (instances). The instance-level labels are unknown.

It is common to classify the whole-slide images into having benign or malignant tissue [39] or try to predict the subtype of cancer [55, 117, 124, 176, 186]. The former can be represented as a binary classification problem, while the latter scenario has multiple options. If the cancer subtypes are mutually exclusive (only one label is possible per whole-slide image, or WSI), then the problem is either a binary classification problem (for two subtypes) or a multi-class classification problem. If multiple classes are possible for a WSI, then each is treated separately in its binary classification problem. The Multiple Instance Learning formulation is easier to explain in the context of a binary classification case with mutually exclusive positive (disease-present) and negative (healthy) classes.

I follow the definition of "healthy" from Section 2.2.2. I only consider the tissue benign if no malignant features suggest otherwise. For a bag (WSI) of instances (patches), it means that if at least one patch contains malignant features, it is sufficient to classify the whole image as having malignant tissue. I can express it mathematically as follows: note that the presentation has been inspired by [117]. Let $B = \{(x_1, y_1), ..., (x_n, y_n)\}$ be the bag of instances x_i and the corresponding labels y_i, where $y_i \in \{0, 1\}$. The label of B, $c(B)$, is given by

$$c(B) = \begin{cases} 0, & \text{iff } \sum y_i = 0 \\ 1, & \text{otherwise} \end{cases} \tag{2.3}$$

Also, MIL formulation uses a suitable transformation f and a permutation-invariant transformation g to predict the label of B:

$$c(B) = g(f(x_1), ..., f(x_n)) \tag{2.4}$$

MIL can be understood in two ways depending on f and g: 1) In the **instance-based approach**, f is an instance classifier that predicts the class for each instance, and g is an aggregation function that combines the instance predictions into a bag prediction. 2) In the **embedding-based approach**, f extracts features (embedding) from each of the instances, g is an aggregation function that combines instance-level features (embeddings) into a bag embedding and converts the bag embedding into the bag prediction. The instance-based method produces the bag predictions via an aggregation of instance predictions, while the embedding-based method's predictions are based directly on bag embedding. This means the embedding-based method is more end-to-end and should be easier to supervise directly from the bag labels as long as the extracted instance-level features are relevant. However, it is easier to determine the key instances from the instance-based methods since we have access to explicit instance-level predictions.

2.3.3 Extracting Features from Patches

From Section 2.3.2, it is known that in both instance-level and embedding-based approaches, it is needed to initially deal with patches of images. In the embedding-based approach, features $f(x_i)$ are extracted from the patch input x_i. In the instance-based approach, f is a classifier; however, a classifier is a feature extractor followed by a classification head. Hence, the patch-level feature extractor plays a crucial role in the overall bag prediction for both approaches.

Researchers can take different approaches for obtaining the feature extractor: using ImageNet pre-trained feature extractor directly [124], fine-tuning ImageNet pre-trained feature extractor using WSI-level labels [176], training the extractor end-to-end using only WSI-level labels on a massive dataset [186], pre-training the extractor using self-supervised methods [24, 40, 46, 73, 117, 125, 136, 187, 189, 188, 194], or on auxiliary tasks with supervised labels [79, 97, 125], training the extractor end-to-end from patch-level annotations when those are available [20, 190, 205].

All of the works mentioned in the previous paragraph presented promising results. The decision about which approaches to use depends on the specific task and the ability to collect more annotated data for your specific task.

2.3.4 Pathology-specific Feature Extractors and Foundation Models

In 2020 and 2021, there was a rise in methods for training computer vision models in a self-supervised manner (see Section 2.1.3). So, in 2021, different computational research groups started using those methods to train feature extractors on pathology datasets. Li et al. [117] utilised pathology-specific feature extractors trained on TCGA-lung [4] and CAMELYON16 datasets [122] using SimCLR [49]. Wang et al. [186] released TransPath, the first pathology model that was trained on multiple cancer classes from TCGA [4] and PAIP [109] datasets, followed by the improved and extended version CTransPath [187].

However, it was not until 2023 and 2024 that the number of tile-level pathology foundation models truly increased, with new models coming almost every month. University research labs released 3B-CPath [40], UNI v1 [46], CHIEF [188], PathDINO [19], and many others; while industry players released REMEDIS [24], Phikon v1 and v2 [73], Virchow v1 and v2 [185], Hibou -b and -L [136], Prov-GigaPath [194], and counting. DINO [41] and DINOv2 [141] have been the most popular methods for self-supervised pre-training of pure vision models.

Inspired by the success of VirTex [64] and CLIP [149] for pre-training natural image feature extractors aligned with text captions (see Section 2.1.3), the computational pathology community released the ARCH dataset [78] (for VirTex pre-training), PLIP [97] pre-trained on OpenPath [96], CONCH v1 [125], CONCH v1.5 [67], and more.

Some of the groups also released slide-level feature extractor models trained on top of their tile-level extractors: Prov-GigaPath [194] tile and slide encoder released together, PRISM [160] slide encoder based on Virchow v1 tile encoder[185], CHIEF [188] tile and slide encoders released together, TANGLE [104] slide encoder on top of UNI [46]

features, TITAN [67] slide encoder released together with CONCH v1.5 [67] tile encoder.

Several benchmarking studies have been released assessing the performance of different foundation models on in-house datasets [36, 48, 77, 69, 137, 191]. All studies agree upon the fact that pathology foundation models improve the downstream performance compared to feature extractors pre-trained on natural images. Breen et al. [36] and Chen et al. [48] agree that while there is no best foundation model for all tasks, AB-MIL remains a strong baseline feature aggregator if used in combination with a suitable feature extractor. The absence of a universally better foundation model to use on your data can be attributed to the varying degrees of similarity between the foundation model training data and your dataset.

An up-to-date list of the available foundation models is being maintained by Georg Wolflein[2].

2.3.5 ROI-level Training on Lung Cancer Images in 2018 and 2019

In 2018 and 2019, the in-house computational pathology datasets collected by research groups were much smaller than now. This fact encouraged research groups to collaborate with pathologists to obtain more fine-grained labels for parts of the WSIs.

Alsubaie et al. [20] and Wei et al. [190] chose Region-of-Interest (ROI) annotations for training. Some of the key steps taken by the works were exactly the same. Both works focused on the adenocarcinoma subtype of lung cancer with the primary focus on five adenocarcinoma patterns (see Section 2.2.2), they included every other pattern in a separate "other" group. They asked pathologists to select ROIs with one out of the (5+1) predominant patterns to get the labels. A sliding-window approach was used to move the (224x224 pixels) focus patch over the ROI and feed it into ResNet18. ResNet's output for a focus patch was the probability distribution over the predominant patterns on this patch. Each focus patch in the ROI was said to have the same predominant pattern as the ROI it came from. The patch-level predictions were aggregated into WSI-level predom-

[2] https://github.com/georg-wolflein/pathology-foundation-models

inant and minor patterns using a heuristic approach developed in collaboration with the pathologists.

The works, however, also had several important differences. The first one is that Alsubaie et al. [20] tried to use different magnification scales and even used early-fusion, *i.e.* concatenated patches at different magnifications to be used as inputs. Also, Alsubaie et al. [20] validated their algorithms on the same type of data, while Wei et al. [190] asked the pathologists to select 224x224 patches with classic examples for each of the (5+1) patterns. Wei et al. [190] tested their models on Whole-Slide-level labels (Section 2.3.6).

2.3.6 WSI-level Training on Lung Cancer Images in 2020 and 2021

I now present two works released in 2021 that successfully used WSI-level labels to train models on lung cancer images and that have significantly influenced my understanding of the field. Li et al. [117] and Lu et al. [124] focused on distinguishing between two most common sybtypes of lung cancer (Lung Adenocarcinoma and Lung Squamous Cell Carcinoma), but did not look into specific morphological patterns. They used the lung portion of the TCGA Dataset [4] with whole-slide-level labels. A sliding window approach was used to generate non-overlapping focus patches (224x224 pixels) over the WSI to be fed into a frozen ResNet feature extractor, resulting in a feature vector summarising the information from every focus patch. Finally, both works used some form of attention mechanism to fuse the feature vectors from different patches into a global WSI diagnosis.

It is important to also highlight the differences between these works. First, Li et al. [117] used non-overlapping patches (stride of the sliding window equals the patch size) while Lu et al. [124] used overlapping patches "to achieve better results on their datasets" at higher computational costs. Second, Li et al. [117] used one multi-scale hybrid branch of attention for all output classes, while Lu et al. [124] used a separate single-magnification attention branch per each output class to select how much evidence each patch brought to

each class. Moreover, Li et al. [117] pre-trained two ResNet18 feature extractors using SimCLR [49] self-supervised learning framework on patches extracted at 2.5x and 10x magnifications [3], while Lu et al. [124] used a truncated version of ResNet50 pre-trained on ImageNet to extract features. Finally, Li et al. [117] trained the model with standard Cross-Entropy loss, while Lu et al. [124] clustered the patches into most and least-evident patches for each class as positive and negative evidence for the class. SVM margin loss was used to ensure that patches bringing positive and negative evidence for each class were clustered separately. Different strategies were used for mutually exclusive and non-mutually exclusive classes.

2.3.7 Comparison of ROI- and WSI-level Training Paradigms

Both ROI- and WSI-level annotations result in a high number of patches per pathologist-given label due to the sliding window approach. *The number of patches per label is greater for the WSI-level approach.*

For ROI-level annotations, the morphological patterns can be explicitly given so the model is forced to look for the patterns preferred by the pathologists. In contrast, the WSI-level approaches use a much coarser classification of labels since small features can not be annotated in this fashion. This means that *networks trained with ROI-level annotations will be inherently better at seeing pathologist-preferred features and thus trusted more by the pathologists.*

2.3.8 Evolution of Aggregation Strategies: 2018 - 2024

I now summarize and compare various patch-aggregation strategies researchers use during the inference stage while working with different cancers. Section 2.3.2 provided a brief introduction to instance-based and embedding-based approaches.

[3]Correction from 5x/20x in the paper: `https://github.com/binli123/dsmil-wsi/issues/21`

Wei et al. [190] and Zhu et al. [205] trained the feature extractors to predict patch-level labels and then used a **heuristic-based aggregation** strategy developed together with the pathologists. They considered the patches with predicted class probabilities higher than a certain threshold to be confidently predicted and used their proportions to determine the overall class. This kind of aggregation is only possible if a patch-level classifier can be trained on patch-level annotations in an instance-based MIL approach.

One of the popular ways to aggregate the patch-level features in an embedding-based approach is to re-weigh the patch feature embeddings using some kind of **attention mechanism** similar to the one introduced in [25]. This allows for end-to-end training when only WSI-level labels are available. There are different entities on a WSI that models can pay attention to: features extracted from image patches, spatial positions within the patches, spatial positions of patches within the WSIs relative to the other patches, and features of the patches within larger patches in the WSIs. Li et al. [117], Lu et al. [124], Tomita et al. [176] used patch-level attention, *i.e.* features extracted from the patches were re-weighted. Fan et al. [71] used attention that was working both in the feature and the spatial dimensions.

Analogous to the work in NLP, using **transformer** architecture [184] was a natural step forward from using attention to aggregate features extracted from patches. This has already been done in multiple works [118, 135, 162, 186]. Instead of tokens (words) in the NLP context, patch-level features were fed into the transformer architecture.

Adding back the **positional relationships** of patches into the models was also shown to be beneficial. One of the ways to imagine this is to put the features extracted from the WSI patches back into an image-like tensor, where a patch feature vector substitutes a patch. After that, a **CNN** [18], a **Slide Gaph** [126], a **Graph-Attention-CNN** [45], or a **ViT** can be used to aggregate the patch-features into a WSI prediction. The convolution operation performed on features from different adjacent patches can capture the local interactions between the morphology of adjacent patches while representing it as a graph, and adding

an attention mechanism allows passing the information about the spatial interactions of the morphological structures on patches between each other.

Finally, it is worth pointing out that pathologists use various **magnifications** when making the diagnosis. Some works use multiple magnifications into their pipelines [20, 18, 117, 128], others try using multiple magnifications and sometimes even combine the predictions in an ensemble model [39, 55], but most of the works I read only use a single magnification [124, 176, 186, 190, 197].

2.4 Gaps and Opportunities

To achieve the goal of connecting pathology and radiology imaging modalities, as described in Section 1.4, required obtaining comprehensive annotations for the pathology slides collected as part of the DART Lung Health Programme. The literature review on Computational Pathology, presented in Section 2.3, highlights that the number of works that mimic the pathologist's workflow - identifying morphological patterns and combining them into a diagnosis - is very limited. In Chapter 3, I address this gap by developing a comprehensive annotation protocol and exploring strategies to efficiently use it for collecting training data in the presence of natural class imbalance.

Another opportunity arises in creating an aggregation mechanism to model the interactions between different morphological pattern labels and cancer subtype labels. The literature review suggests that researchers have been using separate attention branches for predicting multiple labels independently [102, 117, 124] and that no existing works model the interactions between more than two labels or explore how modelling such relationships influences cancer subtyping. This gap is addressed in Chapter 5, where I focus on designing methods that account for interactions between different lung cancer subtypes and adenocarcinoma patterns present in whole-slide images.

Furthermore, to extract better descriptors for morphological features, I leverage patch-

level foundation models described in Section 2.3.4. However, with new foundation models being released almost monthly in 2024, selecting a suitable model to begin with is a challenge. I attempt to address this issue in Chapter 4 by proposing a method to identify a promising foundation model at a fraction of the computational cost compared to exhaustively extracting patch features using all available models and training models on downstream tasks. Although the proposed method is effective for selecting a model for a patch classification downstream task, further research is necessary to facilitate the accurate choice of a foundation model for slide-level downstream tasks.

These gaps and opportunities directly influence the research design of this thesis by guiding the development of efficient annotation protocols, robust aggregation mechanisms that model multiple labels, and computationally feasible strategies for leveraging foundation models. Collectively, these contributions aim to improve lung cancer subtyping and integrate pathology and radiology imaging modalities for improved cancer diagnosis. Due to the issues encountered with the collection and linking CT data described in Section 1.4, I focused primarily on the former aim of improving the subtyping of lung cancers.

Chapter 3

Active Data Enrichment

Contents

Deep learning has revolutionised the computational assessment of digital pathology images. Today, mature algorithms have been developed to assess morphological features at the cellular and tissue levels. In addition, promising efforts are underway to link morphological features with biologically relevant information. While promising, these efforts primarily focus on narrow, well-defined questions.

This work aims to link pathology with radiology with the goal of improving the early detection of lung cancer. Rather than utilising a set of predefined radiomics features, I propose to learn a new set of features from histology. Generating a comprehensive lung histology report is the first vital step toward this goal. Developing a comprehensive pathology report in the given setting requires an annotation strategy that captures all clinically relevant patterns specified in the WHO guidelines. Being comprehensive, it takes more time than the narrower (predicting adenocarcinoma patterns on LUAD slides) or shallower (subtyping LUAD and LUSC) approaches and requires optimisations to be made. The need for a comprehensive annotation protocol that optimally utilises pathologist time motivated the contributions of this work.

Contributions

1. Developed a multistage annotation protocol for lung histology images. Annotation Stages: slide subtyping, region selection, and region annotation.
2. Proposed, assessed feasibility and compared approaches to utilise the pathologist annotation time more efficiently by balancing the dataset and mitigating the biases in learning through interventions at each of the annotation stages.
3. Proposed Ranking AUC - a new metric to measure how well samples from classes of interest can be prioritised. Unlike ROC AUC, the proposed Ranking AUC accounts for unknown and variable annotation time limits and can be applied to supervised and one-shot retrieval methods.
4. Demonstrated the opportunities active data enrichment can provide. While it is possible to automatically prioritise slides and pre-selected regions with clinical patterns under-represented in the dataset, automatically selecting regions to annotate remains challenging.
5. Obtained a new lung cancer dataset annotated to a degree that is not readily available in the public domain.
6. Released the code for Ranking Curve creation, AUC calculation, and plotting: `https://github.com/GeorgeBatch/active-data-enrichment`

This chapter includes my first author paper, "Active Data Enrichment by Learning What to Annotate in Digital Pathology" [31] presented at the Cancer Prevention Through Early Detection (CaPTion) workshop of the Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2022 conference.

3.1 Introduction

The goal of automatically generating a comprehensive pathology report motivates finding new ways to link pathology and radiology in the context of lung cancer. Given the current state of the art in computational pathology, this is an open problem.

Lung cancer accounts for more deaths than any other type of cancer [177]. The three main subtypes are Non-small Cell Lung Carcinoma (NSCC or NSCLC), Small Cell Carcinoma (SmCC), and Carcinoid Tumour. NSCLC accounts for more than 80% of all lung cancer cases [60, 95] and is split into two main subtypes: lung adenocarcinoma (around 50% of all cases [133]) and lung squamous cell carcinoma. Based on existing clinical guidelines, CT is used to detect the presence of lung cancer. Tissue samples are then taken from suspicious regions to confirm the diagnosis. The general type, subtype, and the underlying morphological characteristics of lung cancer determine the clinical prognosis [140]. Hence, it is vital to identify all subtypes, including those that occur less frequently, in a robust manner. The difficulty in making an accurate diagnosis lies in the inter- and intra-tumour heterogeneity [203]. The large inter-observer variability in tumour subtyping [165] is another factor that needs to be taken into account.

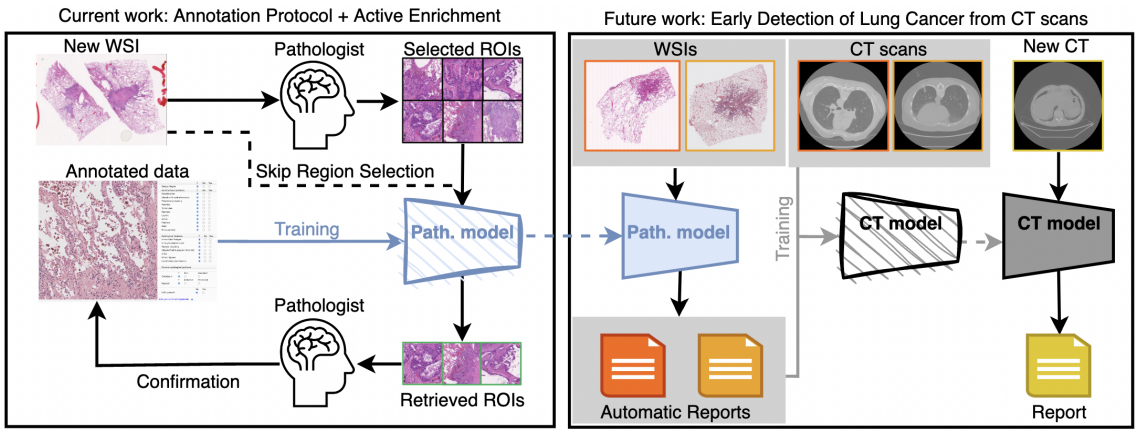

Figure 3.1: Annotation protocol and active data enrichment (left) in the context of early detection of lung cancer from CT images (right). A trained pathology model will generate automatic histology reports for new WSIs. The generated reports will be used together with corresponding chest CT scans to learn a new set of radiology features in order to improve the early detection of lung cancer. Models in training are shown with sketch-style filling, while solid fill represents trained models during the inference stage.

Recently, a number of promising approaches for the automatic subtyping of specific lung cancers and identifying specific lung cancer morphologies have been published. With the help of extensive manual supervision in the form of region-based annotation, it is now possible to determine the predominant morphological pattern of lung adenocarcinoma [20, 190]. Other works used WSI-level labels for weak supervision. Binary subtyping of NSCLC into adenocarcinoma and squamous cell carcinoma has been performed in several works [55, 117, 124]. Yang et al. [197] extended it to six different disease types by adding samples from small cell lung carcinoma, pulmonary tuberculosis, organising pneumonia, and normal lung tissue.

The drawback of all these methods is that they either identify the morphologies of adenocarcinoma (the most prominent type of lung cancer) and do not take other types of lung cancer into account [20, 190] or classify lung cancer types directly from the histology images, omitting the stage of explicitly finding the morphological features used by pathologists to make the diagnosis [55, 117, 124, 197]. All referenced works in the latter group present the WSIs with overlaying heatmaps to show that trained models pay attention to clinically relevant regions; however, it has been shown that when using different interpretability methods, "reliance, solely, on visual assessment can be misleading" [17].

To support our goal of identifying new features that aid the early detection of lung cancer on CT, we require an approach that closely mimics how pathologists work today. It is critical to automatically identify and aggregate a broad range of WHO-defined features [140] at different magnifications (Sections 2.2.2 and 3.3) to make the final diagnosis. To this end, I developed a novel annotation protocol (Figures 3.2, 3.3, 3.4), which makes optimal use of the available data and expert annotation time. To utilise the limited time human experts can dedicate to such an annotation task, I developed an approach that actively selects specific cases to achieve a balanced training dataset. This work focuses on discussing and analysing novel techniques to optimise the annotation process (Figure 3.1, left).

Coarse labels can include subtyping, which can be extracted from the pathology reports if available. Finer annotation can include exhaustive or non-exhaustive selections of diagnostically relevant tissue, which in our case will likely be tumour tissue. To achieve an even finer annotation level, each region's morphological characteristics can be recorded.

The non-uniform and complex distribution of lung cancer subtypes means that one will collect many samples of the most common subtype (lung adenocarcinoma) and fewer samples of the rest of the subtypes. Each subtype is characterised by its own morphological patterns, which means that the characteristic patterns of under-represented subtypes will be under-represented in the dataset. However, when the samples come without annotations, one can aim to mitigate the bias towards the most common class and facilitate the learning of all classes by prioritising the annotation of subtypes and patterns for which few labels are available. Similar to the three levels of labels described in the previous paragraph, it is possible to intervene and influence which slides should be annotated, which regions on slides should be selected for annotation, or, if regions have already been pre-selected, which of them should be annotated.

3.2 Dataset and Annotation Protocol

In this section, I describe the original annotation protocol I used when only 37 digitised slides were available, the updated annotation protocol I employed after the number of slides started rapidly increasing, and the dataset received by the end of the DART lung health project, which was based on top of the Targeted Lung Health Checks programme run by NHS England. During the project, an in-house digital slide viewer AIDA [14] was set up to present the digitised slides to the pathologists.

3.2.1 Limited Data Setting

Stage 1: Region Selection.

I asked the pathologists to annotate enough diagnostically relevant regions (Regions of Interest, ROIs) on each slide to support a diagnosis (Figure 3.2). Given that only 3–5% of cases in lung cancer screenings are expected to be benign, I anticipated a similarly small proportion of received slides to contain no cancer tissue. However, to help the models learn how non-cancerous tissue appears—and how it differs from cancerous tissue—I needed additional examples. Therefore, I asked the pathologists to select one or two non-cancerous regions on slides that also contained cancer. According to the DART pathologists, when shown in isolation, these regions are indistinguishable from those on entirely benign slides, making them valuable for training models to recognise non-cancerous morphology even in the presence of cancer elsewhere on the slide. To ensure a comprehensive representation of tissue morphology, pathologists annotated regions at different magnifications: lower magnification to capture *architectural patterns* (green ROIs), and higher magnification to highlight *cytological features* (yellow ROIs). It was not necessary for high-magnification regions to be spatially nested within low-magnification regions, as the two scales of information capture complementary aspects of diagnosis and are not required to always be contextually linked in the same region. The pathologists selected 7–14 regions per slide (10 regions selected on average), spending 5–10 minutes per slide.

Stage 2: Region Annotation.

The aim of this step is to use the terms from the 2021 WHO Classification of Lung Tumours [140] to annotate each of the ROIs that have been selected at the previous stage. All relevant labels are shown on the right of Figure 3.3. The "?" label is introduced to mitigate inter-observer variability. Only unequivocal cases are given definite labels. Additionally, the annotation platform presents a question of whether an EVG-stained patch would be useful in order to make a diagnosis. I did not implement automatically showing the EVG-

stained patch, because subsequent tissue slices would have required alignment. However, recording the desirability of the EVG-stained slide can be used to develop software for automatically requesting EVG-stained slides, similar to the automatic IHC requests proposed by Chatrian et al. [42] for prostate cancer biopsies. Answering all 22 questions for each ROI from a slide with 10 selected ROIs took around 20 minutes (2 minutes per region), making it by far the most time-consuming part of the annotation process.

3.2.2 Abundant Data Setting

When I started receiving more whole-slide images than the pathologists could annotate following the detailed annotation protocol, I modified the annotation protocol in three ways to make it more efficient.

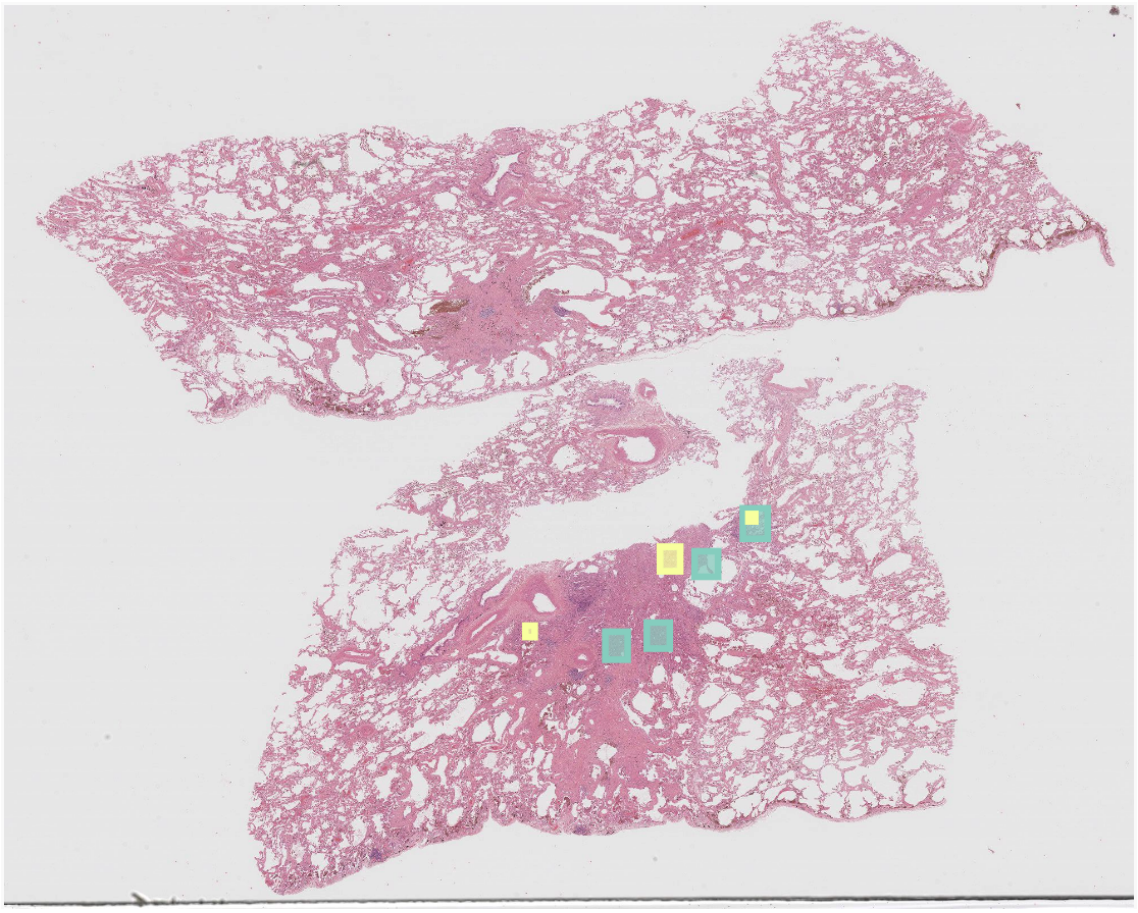

Figure 3.2: **Annotation Stage 1: Region Selection.** A pathologist chooses a sufficient number of relevant regions of interest at different magnifications to support a diagnosis.

First, I added a **subtyping stage** to capture the main types of lung cancer described in the WHO guidelines [140] for each slide (Figure 3.4) without marking the diagnostic locations and labelling morphological patterns on them. Note that pathologists could select multiple cancer subtypes for a single slide (I received a slide with both non-mucinous lung adenocarcinoma and squamous cell carcinoma tissue). Unlike the region selection stage, which takes 5-10 minutes per slide, or the region annotation stage (20-30 minutes per slide, depending on the number of selected ROIs), the subtyping stage takes 2-3 minutes, which is comparable to answering 22 questions for a single selected region from the limited data setting protocol (Section 3.2.1). The goal was to have the pathologists subtype all incoming whole-slide images from the DART project.

Second, I began **collecting the region annotation information earlier**, during the region selection stage. In the original annotation protocol, I asked the pathologists to choose

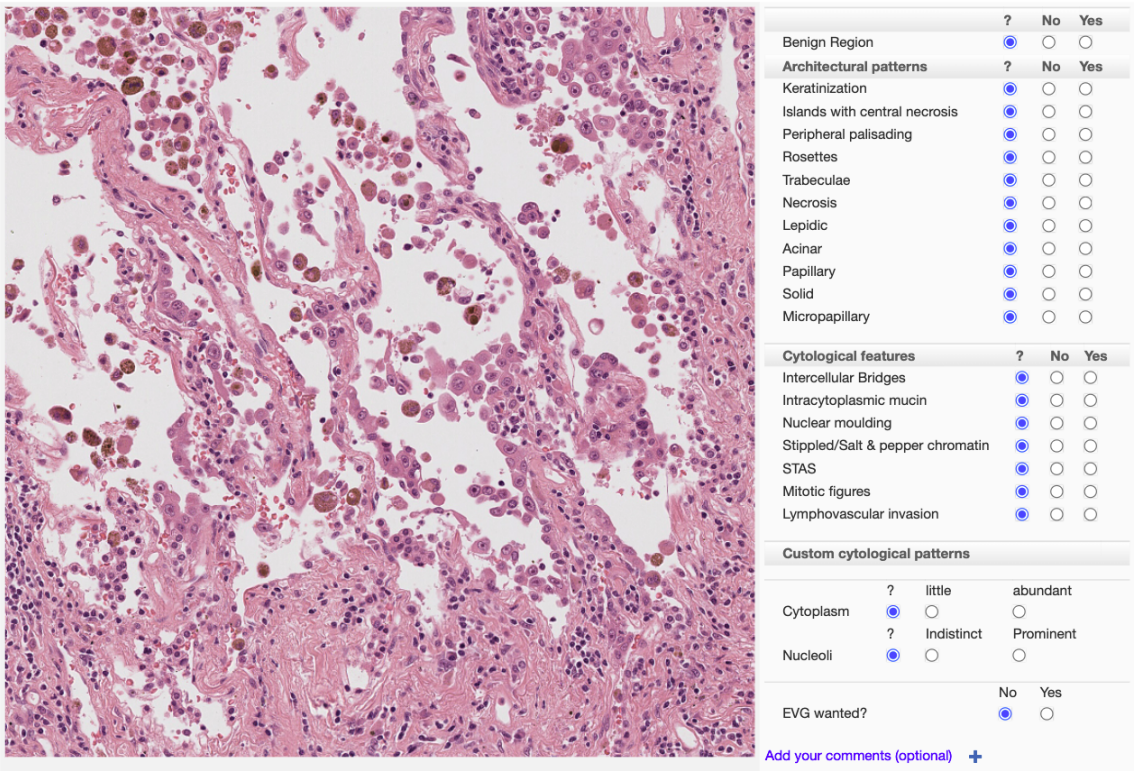

Figure 3.3: **Annotation Stage 2: Region annotation.** View for one of the ROIs. A pathologist is asked to identify whether the region is benign, mark the presence and absence of architectural patterns and cytological features, and indicate the desirability of an EVG-stained version of the region.

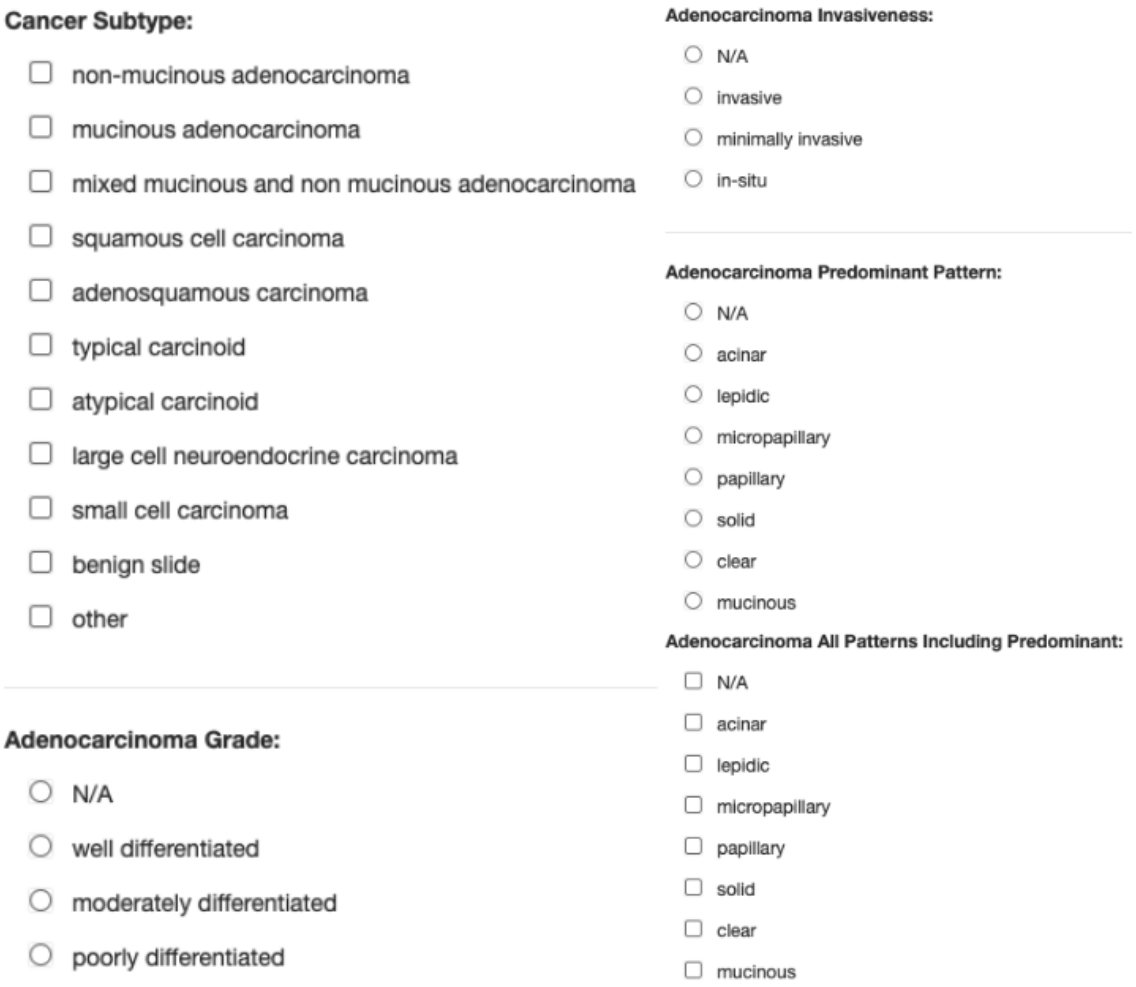

Figure 3.4: **Slide Subtyping Annotation Stage.** Pathologists are asked to select the cancer subtype(s) present on the slide, as well as grade, invasiveness, predominant and other patterns for adenocarcinomas.

representative regions at different magnifications, saving "Low Power" or "High Power" as region metadata, which proved to be not useful. In a new annotation protocol, I asked the pathologists to choose "Diagnostic Regions" and "Benign Regions", saving the status as region metadata. Note, on slides with only benign tissue, no regions needed to be selected, as all tissue could be classified as benign from the slide label. This change to the annotation protocol provided useful information about the regions without requiring extra time and enabled direct supervision of models by highlighting the regions they should and should not focus on while selecting the cancer subtype on the presented slides.

Finally, after receiving even more slides, I introduced a **partial region selection** procedure into the subtyping stage. I asked the pathologists to select only the diagnostic regions that they used to make a decision during the subtyping stage, while still asking them to select a few benign regions. This change increased the amount of information from the subtyping stage, with no extra time needed from the pathologists.

Since the region selection process in our annotation protocol is non-exhaustive, meaning that large portions of tissue on each slide remain unlabelled, it is crucial to include both diagnostic and benign regions if the data is to be used for supervising MIL models (Section 2.3.2) as done in Chapter 5. If only diagnostic regions are selected, region-based supervision introduces two key limitations: (1) benign slides, containing no diagnostic regions, would present difficulties at inference time, and (2) diagnostic slides would lack examples of non-diagnostic tissue, leading to degenerate models that trivially assign diagnostic status to all patches. Including both types of regions allows the model to learn meaningful distinctions between cancerous and non-cancerous tissue, thereby improving generalisation in mixed-supervision settings where both slide-level and non-exhaustive region-level labels are available.

3.2.3 DART Dataset

In this section, we present an overview of the data collected from Oxford University Hospitals. The annotation protocol evolved in response to the changing pace and scale of data collection during the DART project. Early on, before most of the slides had been received, we obtained an initial batch of 37 H&E slides from Oxford University Hospitals (OUH). With annotation time available, the pathologists were able to follow the detailed protocol described in Section 3.2.1. These early slides were annotated more thoroughly than those that followed and helped establish a reference for the kinds of diagnostic patterns we aimed to capture.

As the project moved forward, the remaining 175 slides from OUH arrived, along with additional slides from other DART hospitals. The increasing volume and the limited time left in the project made it impractical to continue with the detailed protocol for all incoming slides. To ensure broader dataset coverage while making efficient use of the pathologists' time, we introduced a faster annotation protocol (Section 3.2.2), which focused on capturing key diagnostic regions more quickly.

Another factor influencing this shift was the availability of pathology reports. While OUH reports could be accessed to support detailed annotations, reports from other hospitals could not be shared due to data governance restrictions: personal information could not be automatically redacted from scanned documents. Taken together, these factors led to a practical adjustment in strategy, prioritising wider annotation coverage over per-slide detail in the later stages of the project. A full timeline of DART data collection is provided in Section 1.4.

Figure 3.5 shows the label distribution for all 265 annotated regions from the first 37 slides received from OUH. In contrast, Table 3.1 shows the summary of all annotations received as part of the DART project. Columns "Patients", "Slides", and "Slides Subtyped" of Table 3.1 show that for the DART sites 1-4, we were receiving many more slides per patient than from OUH. Having limits on annotation time, I asked the pathologists to annotate only the

	no	yes	not sure	total	
Benign region	234	13	18	265	
	no	yes	not sure	total	
Keratinization	238	16	11	265	
Islands with central necrosis	251	9	5	265	
Peripheral palisading	257	2	6	265	
Rosettes	236	23	6	265	
Trabeculae	248	12	5	265	
Necrosis	237	20	8	265	
Lepidic	165	64	36	265	
Acinar	144	54	67	265	
Papillary	255	3	7	265	
Solid	199	60	6	265	
Micropapillary	199	29	37	265	
Intercellular Bridges	157	9	99	265	
Intracytoplasmic mucin	199	5	61	265	
Nuclear moulding	213	0	52	265	
Stippled/Salt & pepper chromatin	121	45	99	265	
STAS	253	1	11	265	
	no	yes	total		
EVG wanted	197	68	265		
	little	abundant	not sure	total	
Cytoplasm	0	217	48	265	
	indistinct	prominent	not sure	total	
Nucleoli	80	82	103	265	
	< 2	2 - 10	> 10	not sure	total
Mitotic count	12	0	0	193	205

Figure 3.5: **Label distribution for all annotated regions** that were pre-selected by the pathologists for the first two batches (37 slides) from Oxford University Hospitals (OUH).

most relevant 1-2 slides per patient. Although this does not change the eventual number of slides that pathologists can annotate within a given time interval, annotating slides from more patients increases the diversity of annotated slides compared to annotating the same number of slides from fewer patients. Increasing the data diversity is important for training machine learning models since it makes models more generalisable. It is also important for the evaluation stage since it makes the evaluation closer resemble the diversity of the real-world data.

Site	Patients	Slides	Slides Subtyped	Slides with ROI Selection	ROI Annotations
OUH 1	15	20	20	20	145
OUH 2	15	17	17	17	120
OUH 3	159	175	175	142	-
DART 1	26	58	58	53 (34 + 19)	-
DART 2	42	120	69	48 (44 + 4)	-
DART 3	91	290	86	86 (86 + 0)	-
DART 4	32	124	72	72 (71 + 1)	-
Total	380	804	497	438 (414 + 24)	265

Table 3.1: **In-house data summary.** Slides from Oxford University Hospitals (OUH) came in 3 batches. Locations of DART sites 1-4 were coded to increase data security. Each patient had at least one H&E slide, which could be a biopsy or a resection. The annotation protocol described in Section 3.2.2 was employed to subtype slides. Some of the subtyped DART slides did not have enough tissue to determine the cancer type and patterns present. Region of Interest (ROI) selection was performed in a complete setting (see Section 3.2.1) for some slides and in a partial setting for others (see Section 3.2.2). **Example:** for the site "DART 1", we got regions selected for 53 slides (34 complete and 19 partial selections). 265 ROIs from the first 37 slides we received from the Oxford University Hospitals (OUH) have been annotated as described in Section 3.2.1.

Site	LUAD	LUSC	TC	Other Cancer	Benign	Total
OUH 1	12	0	5	1	2	20
OUH 2	6	7	4	0	0	17
OUH 3	120	17	33	4	2	175†
DART 1	23	4	4	1	3	35
DART 2	26	12	3	11	9	61
DART 3	36	12	0	1	16	65
DART 4	36	5	1	3	11	56
Total	258	56	50	21	43	428†

Table 3.2: **Cancer subtype distribution of labelled data.** Locations of DART sites 1-4 were coded to increase data security. The annotation protocol described in Section 3.2.2 was employed for subtyping (Table 3.1, column "Slides Subtyped") slides. LUAD - lung adenocarcinoma, LUSC - lung squamous cell carcinoma, TC - typical carcinoid, Other Cancer - minority subtypes, and Benign - non-cancerous tissue. † If you sum up the values in the row corresponding to the 3rd OUH batch (OUH 3), you will get 176 instead of 175. The reason for this is that one slide had both non-mucinous adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC) present on it.

Site	Acinar	Lepidic	Micropapillary	Papillary	Solid
OUH 1	11 (2)	11 (8)	7 (1)	0 (0)	2 (1)
OUH 2	6 (2)	6 (3)	4 (1)	1 (0)	3 (0)
OUH 3	114 (43)	106 (44)	87 (13)	37 (4)	35 (14)
DART 1	19 (8)	11 (8)	10 (3)	6 (1)	8 (3)
DART 2	18 (8)	11 (6)	15 (4)	4 (4)	12 (4)
DART 3	27 (9)	27 (17)	28 (4)	9 (2)	12 (4)
DART 4	28 (11)	18 (5)	26 (12)	2 (0)	18 (8)
Total	223 (82)	190 (91)	176 (38)	59 (11)	89 (34)

Table 3.3: **Presence (and predominance) of adenocarcinoma patterns** on adenocarcinoma slides (Table 3.2, column "LUAD"). Each adenocarcinoma slide has only one predominant pattern, but multiple adenocarcinoma patterns can be present. **Example:** The 1st OUH batch (OUH 1) has 11 adenocarcinoma slides with an acinar pattern present, but only on 2 of them, the acinar pattern is predominant.

3.3 Methodology

The annotation goal was to obtain slide labels, diagnostic region selections, and region labels at different magnifications to support automated reporting of all clinically relevant subtypes of lung cancers. Here, we take the WHO guidelines [140] as a reference. The labels should include the features and patterns used by the pathologists for making the diagnosis from WSIs. We chose this bottom-up approach of first finding diagnostically relevant regions and only then aggregating the information found in them into a slide-level diagnosis for two reasons: first, to mimic the pathologists' workflow as closely as possible, and second, because slide-level pathology reports were already available, so there was no need to ask the pathologists to do this again.

3.3.1 Dataset Enrichment

Due to the high cost of expert annotation, it is vital to optimise the annotation process. For us, it means minimising the time spent by the pathologist in order to achieve a quality of the data sufficient for efficient model training. When training a model to recognise multiple classes at once, it is crucial to create a balanced dataset in which all classes are well represented.

A naive sequential annotation of the available data would naturally result in an extremely unbalanced dataset in which slides with rare subtypes and patterns of rare disease subtypes would be under-represented. Attempting to get a sufficient number of under-represented subtypes and patterns in a naive sequential manner would result in sub-optimal use of limited expert annotation time. To address this, I propose two approaches to increase the representation of under-represented subtypes and patterns, enabling the models to learn distinguishing features through the application of known image retrieval techniques. The approaches are illustrated in Figure 3.1 (left): ranking regions pre-selected by the pathologist (solid arrows), automatically selecting regions from non-annotated WSIs in a sliding-window sweep (dashed line), and selecting the slides to present for annotation. Note that I am not proposing to bootstrap or resample already annotated images; instead, I am proposing to change the order of annotation from random to systematic, which can help increase the efficiency of annotations by introducing a bias towards annotating samples from the class of interest.

Due to the data availability constraints, I only explored the one-shot retrieval methods for slide prioritisation and automatic ROI selection. The main focus of this work was on ranking regions pre-selected by the pathologists, for which I explored both supervised and one-shot retrieval methods (results in Section 3.4).

One of the limitations of this work is the assumption that for different subtypes and patterns, recognition and enrichment are independent. In reality, groups of classes can coexist. This might cause a problem: *if* (1) class A and class B coexist and are usually seen together; (2) class A is easier to spot than class B for the retrieval method, *then* the retrieval method can learn to predict the presence of class B whenever class A is present and disregard any features specific to pattern B. This would lead to false positive predictions of class B when only class A is present.

This work was a proof-of-concept study showing that it is possible to enrich for certain morphological patterns appearing on pre-selected regions using the proposed methods

if needed. The label distribution for the pre-selected regions at the time of the paper publication is shown in Figure 3.9.

In order to measure the retrieval performance of ranking methods for selecting slides and choosing ROIs from pre-selected regions, we proposed a new metric, Ranking Curve AUC, that we describe in Sections 3.3.2 and 3.3.3.

As a result of this work, limited pathologist annotation time could be utilised with greater efficiency.

3.3.2 Ranking Curve AUC - Intuition

When data collection is ongoing, the total number of annotated samples—and particularly the number of positive samples—may be unknown. To address this, we propose a metric for evaluating retrieval performance when the number of available samples varies: **Ranking Curve AUC**. This metric is conceptually similar to ROC AUC, which is widely used for classification tasks.

For each n, we define the **ranking score** as the proportion of *positive samples* (from the under-represented class we aim to enrich) among the top-n ranked samples, relative to the maximum number of positive samples that could be found in any n samples, i.e., $\min(n, t)$, where t is the total number of positive samples. We use this denominator because it is impossible to retrieve more than t positive samples regardless of n.

Unlike **ROC AUC**, which is threshold-agnostic but assumes a fixed dataset, Ranking Curve AUC is designed for scenarios in which the number of positives is unknown and the dataset size changes dynamically during the annotation process. This makes it more suitable than ROC AUC for our setting.

Alternative ranking metrics, such as the **Cox loss** and the **Concordance Index (c-index)**, could also be considered. Cox loss (or Cox partial likelihood) is the objective function used in fitting Cox proportional hazards models in survival analysis. It evaluates how well

the model predicts the relative ordering of event times, focusing on risk ranking rather than absolute outcomes. Lower values indicate a better fit. The Concordance Index, a generalisation of ROC AUC for survival data, accounts for censoring and measures the proportion of correctly ordered, usable subject pairs based on predicted risk scores.

However, both Cox loss and the Concordance Index are designed for continuous-time outcomes and are less suited to binary classification tasks, such as determining whether a region of interest (ROI) belongs to a specific class. They are also sensitive to arbitrary mis-orderings of tied positive or negative pairs. Moreover, neither metric accounts for the dynamic nature of the annotation process, where both dataset size and class balance evolve over time. These limitations make **Ranking Curve AUC** a more appropriate metric for our use case.

Figure 3.6 illustrates examples of one-shot retrieval and supervised ranking procedures and the corresponding ranking curves. For one-shot retrieval (left), let *query* be a sample already annotated to have the class of interest. We can use a distance metric of choice to prioritise positive samples (green circles) based on their distance from the query sample (green star) through one-shot retrieval: the closer the sample, the higher the rank. Standard distance metrics would require samples to be represented by a numeric vector, or feature vector, in order to calculate the distance between samples. For supervised retrieval (right), the predicted probability of coming from the positive class is used to prioritise positive samples (blue circles): the higher the probability, the higher the rank. For this method, we need a classifier that was trained on the samples for which we already have labels.

To calculate the AUC, we can connect discrete data points of the ranking curve with straight lines. To facilitate comparison of retrieval strategies if the total number of samples changes, I normalise the Ranking AUC to have an upper bound of 1 when all relevant samples are ranked higher than irrelevant ones. For a formal definition and the derivation of properties, see Section 3.3.3.

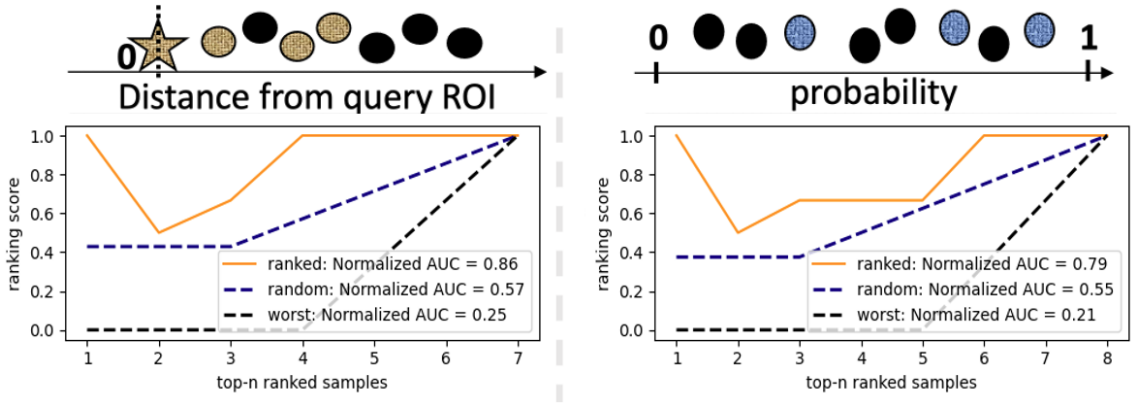

Figure 3.6: Retrieval strategies. Textured dots represent ROIs with the patterns of interest. **Left: One-shot retrieval**. ROIs are ranked in increasing order of distance to the query ROI, an annotated ROI with the pattern of interest. The distance can be computed with a metric of choice using the feature vectors extracted from the ROIs with a pre-trained feature extractor. **Right: Supervised**. For the ROIs, a trained classifier predicts the probabilities of having the pattern of interest present, which are used to rank the ROIs.

3.3.3 Ranking Curve AUC - Mathematical Formulation

Ranking Curve Definition

- Let inputs $\{x_1, ..., x_N\}$ be already ranked by some ranking method that aims to put the inputs with the positive label (1) before (at lower indexes) the inputs with the negative label (0).
- Let $\{y_1, ..., y_N\}$ such that $y_i \in \{0, 1\}$ for all $i \in \{1, ..., N\}$ be the true labels
- Let t be the total number of positive labels
- Let $p = t/N$ be the proportion of positive samples in our data.

$$t = \sum_{1}^{N} y_i \tag{3.1}$$

We define the **ranking score** $R(n)$ as follows for each $n \in \{1, ..., N\}$

$$R(n) = \frac{\sum_1^n y_i}{min\{n, t\}} \tag{3.2}$$

- The numerator is the number of positives among the top-n ranked samples: $\sum_1^n y_i$
- The denominator is the total number of positive samples that we could have possibly retrieved by taking n samples, which is $min\{n, t\}$ since we can neither retrieve more positive samples than there are in total (t), nor can we retrieve more samples than we took (n)

Expected Performance on a Random Sample

If the samples are not in any particular order, the positive samples will be uniformly distributed within the indexes $\{1, ..., N\}$, giving each index a probability $p = t/N$ of having a positive sample. This means that for any sample of n randomly chosen samples, we have:

$$\mathbf{E}[\sum_{1}^{n} y_i] = p \times n \tag{3.3}$$

Which results in:

$$\begin{aligned}
\mathbf{E}[R(n)] &= \mathbf{E}[\frac{\sum_1^n y_i}{min(n,t)}] \\
&= \frac{p \times n}{min\{n,t\}} \\
&= \frac{(t/N) \times n}{min\{n,t\}} \\
&= \begin{cases} t/N, & 1 \leq n \leq t \\ n/N, & t < n \leq N \end{cases} \\
&= \begin{cases} p, & 1 \leq n \leq t \\ n/N, & t < n \leq N \end{cases}
\end{aligned} \tag{3.4}$$

Plotted as a function of n from 1 to N, this results in a horizontal line for $1 \leq n \leq t$ and

a line with $1/N$ gradient for $t \leq n \leq N$. The two lines have the common point when $n = t$, where $\mathbf{E}[R(n)] = t/N = p$.

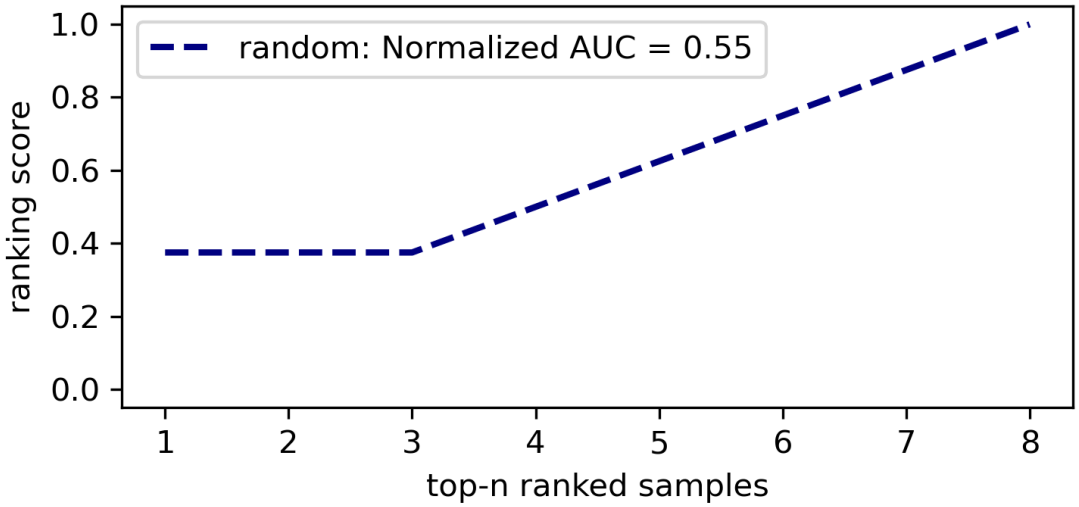

Figure 3.7: Expected Ranking Curve for $N = 8, t = 3, p = 3/8 = 0.375$. This curve represents the expected scenario, in which we have 8 samples (5 negative and 3 positive). Since there is no ranking involved, positive samples are equally likely to be in any of the positions from 1 to 8.

$$AUC = (1 \times (N-1))/2 + ((t-1) \times p)/2 \tag{3.5}$$

- $(N-1)/2$ is the area of the right triangle between (1, 0), (N, 0), (N, 1).
- $((t-1) \times p)/2$ is the area of the right triangle between (1, 0), (1, p), (t, p).

To calculate the Normalised AUC we change the x-axis from $[1, N]$ to $[0, 1]$ which results in

$$AUC = \frac{(1 \times (N-1))/2 + ((t-1) \times p)/2}{N-1} = 0.5 + \frac{t-1}{N-1} \times \frac{p}{2}, \tag{3.6}$$

For the example on Figure 3.7, Normalised AUC $= 0.5 + (2/7) * (3/8)/2 \simeq 0.55$.

Lower Bound

In the worst case scenario, all positive samples are ranked after all negative samples, i.e. $y_i = 0$ for $i \in \{1, ..., N-t\}$ and $y_i = 1$ for $i \in \{N-t+1, ..., N\}$, the Normalised AUC = $(t/2)/(N-1)$ because:

$$
\begin{aligned}
R(n) &= \frac{\sum_1^n y_i}{min(n,t)} \\
&= \begin{cases} 0, & 1 \leq n \leq N-t \\ n-(N-t), & N-t+1 < n \leq N \end{cases}
\end{aligned} \tag{3.7}
$$

This forms a right triangle between (N-t, 0), (N, 0), and (N, 1) with an area of $(t \times 1)/2 = t/2$. When normalising the x-axis from [1, N] to [0, 1], the area becomes $(t/2)/(N-1)$. Substituting $t = p*N$ gives another formula for Normalised AUC = $(p/2)*(N/(N-1))$.

In the worst case, shown on Figure 3.8, the Normalised AUC $= (3/2)/(8-1) \simeq 0.21$.

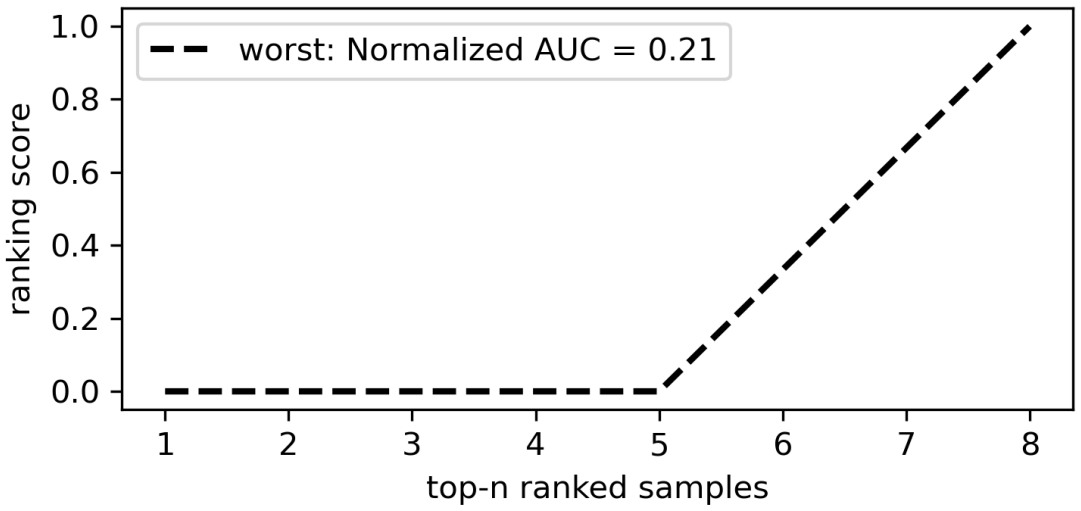

Figure 3.8: Minimal Ranking Curve for $N = 8, t = 3, p = 3/8 = 0.375$. This curve represents the worst-case scenario, in which we have 8 samples (5 negative and 3 positive), and we rank the 3 positive samples last, placing them at positions 6, 7, and 8.

Upper Bound is 1

In the best case scenario, all the positive samples are ranked before all negative samples, i.e. $y_i = 1$ for $i \in \{1, ..., t\}$ and $y_i = 0$ for $i \in \{t+1, ..., N\}$, the Normalised AUC = 1 because for each $n \in \{1, ..., N\}$ $R(n)$ equals 1:

$$\begin{aligned} R(n) &= \frac{\sum_1^n y_i}{min(n,t)} \\ &= \begin{cases} n/n, & 1 \leq n \leq t \\ t/t, & t < n \leq N \end{cases} \\ &= \begin{cases} 1, & 1 \leq n \leq t \\ 1, & t < n \leq N \end{cases} \\ &= 1 \end{aligned} \tag{3.8}$$

3.4 Results: Ranking Pre-selected Regions

I compare one-shot and supervised retrieval strategies for ranking regions pre-selected by the pathologist. Figure 3.9 shows the distribution of labels for ROIs extracted from the first 37 images we received from OUH. To study the effectiveness of the enrichment methods, a pathologist annotated 2 batches with 20 and 17 WSIs that resulted in 145 and 120 selected ROIs, respectively (see Table 3.4). The batches were scanned with different scanners at 20x magnification, which could decrease the retrieval performance due to the distribution shift.

For this proof-of-concept work I selected to mine regions with keratinization pattern via one-shot retrieval and the acinar pattern using a supervised model.

	no	yes	?
Benign Region	0.88	0.05	0.07
	no	yes	?
Keratinization	0.90	0.06	0.04
Islands with central necrosis	0.95	0.03	0.02
Peripheral palisading	0.97	0.01	0.02
Rosettes	0.89	0.09	0.02
Trabeculae	0.94	0.05	0.02
Necrosis	0.89	0.08	0.03
Lepidic	0.62	0.24	0.14
Acinar	0.54	0.20	0.25
Papillary	0.96	0.01	0.03
Solid	0.75	0.23	0.02
Micropapillary	0.75	0.11	0.14
Intercellular Bridges	0.59	0.03	0.37
Intracytoplasmic mucin	0.75	0.02	0.23
Nuclear moulding	0.80	0.00	0.20
Stippled/Salt & pepper chromatin	0.46	0.17	0.37
STAS	0.95	0.00	0.04
	no	yes	
EVG wanted?	0.74	0.26	
	little	abundant	?
Cytoplasm	0.00	0.82	0.18
	indistinct	prominent	?
Nucleoli	0.30	0.31	0.39

Figure 3.9: **Label distribution for the pre-selected regions that have already been annotated for batches 1 and 2 at the time of the publication.** All proportions are given from 265 annotated regions. Note that, in both batches combined, the *acinar* pattern is almost as well-represented as the lepidic pattern. However, in the first batch, there were 54 images with the lepidic pattern and only 34 with the acinar pattern; hence, the selection of the acinar pattern is under-represented.

Batch	WSIs	ROIs	Keratinization	Acinar
(1) Hamamatsu, 20x	20	145	1	34
(2) Ventana DP200, 20x	17	120	15	20

Table 3.4: **Dataset distribution of the first two OUH batches of images.** Two patterns were chosen for performing the experiments (one-shot retrieval: keratinization, supervised: acinar).

Feature Extractor	Pre-training	Success Rate	Ranking AUC
Truncated ResNet50 [124]	ImageNet	3/20	0.62
ResNet18 w/o last layer	ImageNet	4/20	0.62
ResNet18 w/o last layer	TCGA patches at 2.5x [117]	7/20	0.73
ResNet18 w/o last layer	TCGA patches at 10x [117]	**8/20**	**0.79**
Random	NA	2.5/20	0.51

Table 3.5: **Retrieval performance of different feature extractors.** The proportion of regions with keratinization in the second batch is $15/120 = 1/8$, meaning that we expect $20 * 1/8 = 2.5$ samples with keratinization in any random sample of 20 samples.

3.4.1 One-shot Retrieval Data Enrichment

Keratinization was the most under-represented in the first OUH batch, with only 1 of the ROIs marked to have it by our pathologist, as illustrated in Table 3.4. Having only one image of a particular class suggested using one-shot image retrieval since a good model cannot be trained from one sample. Furthermore, this region had no other patterns from the annotation list. Hence, I enriched the keratinization class as follows. I used the region as a *query region*. It was passed through a feature extractor (see Section 2.3.3) to create a *query vector*. Not-annotated *target regions* pre-selected by the pathologist were processed by the same feature extractor. Matches with the smallest cosine similarity, a distance metric commonly used for image retrieval, were then presented to the pathologist for confirmation.

15 out of 120 ROIs in the second batch contained the keratinization pattern. I simulated the situation of prioritising which of the 120 patterns should be annotated to maximise the number of images with keratinization in the top-20 ranked samples, since it was the number of ROIs that the pathologist reported being able to annotate in a single annotation session. Since extra annotation was not required, I tried different feature extractors available

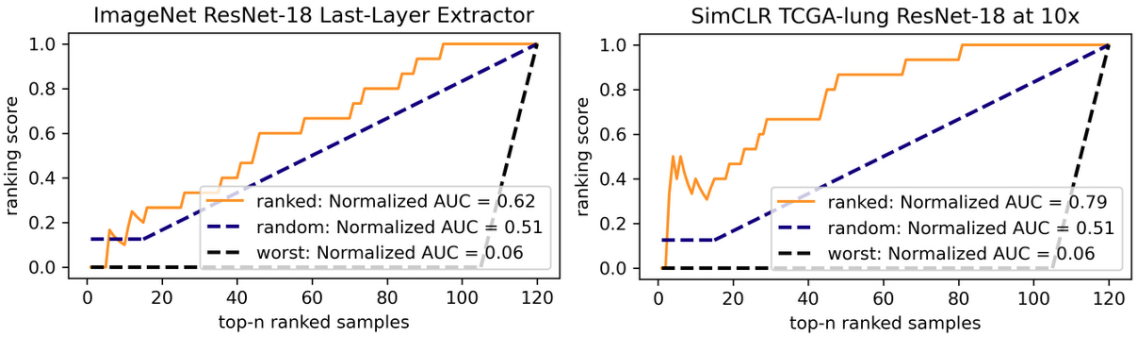

Figure 3.10: Solid orange line: ranking curve (as described in Section 3.3.1). Dashed blue line: expected cumulative proportion if selecting samples at random. ImageNet trained extractor (left) shows worse results than the TCGA-lung pre-trained extractor (right).

at the time: a modified version of an ImageNet pre-trained ResNet18 without the classification layer, an ImageNet pre-trained Truncated ResNet50 [124], and two ResNet18 models pre-trained using self-supervised learning on patches extracted at 2.5x and 10x magnifications from TCGA-lung by Li et al. [117] (results in Table 3.5). ViT [68] foundation models (Section 2.3.4) pre-trained on pathology data were not yet available when this experiment was conducted. The TCGA-lung pre-trained networks performed better than the ImageNet pre-trained ones. However, 20 was still an arbitrary number of ROIs to consider, and this choice could seriously affect the performance evaluation. Hence, I used the Ranking Curve AUC proposed in Section 3.3.1. The curves for two feature extractors are shown in Figure 3.10 while the AUC values are presented in Table 3.5. ResNet18 feature extractor pre-trained on patches of TCGA-lung extracted at 10x magnification [117] showed best retrieval results. For any 40 regions chosen at random, it was expected to have $40/120 \approxeq 0.33$ or 5 out of 15 regions with the keratinization pattern. However, the top-40 ranked regions had 10 out of 15 regions, doubling the proportion compared to random choosing ($10/15 \approxeq 0.66$). Labels for two-thirds of the regions with the keratinization pattern could be obtained with only a third of annotated regions from the second batch.

3.4.2 Supervised Active Data Enrichment

I then explored how supervised enrichment methods could be used to prioritise the regions with the acinar pattern present. I evaluated how including these regions improved the classifier's performance on a previously unseen test set and mitigated the biases in learning by reducing class imbalance.

I split the dataset into 4 subsets: *Train*, *Validation*, *Pool*, and *Test,* with 20/86, 14/59, 10/60, and 10/60 samples containing the acinar pattern, respectively. Table 3.6 shows the detailed data distribution of acinar patterns in different sets. The pool set imitates the regions next in line for annotation - they are not annotated yet, but there is an option to annotate some of them. The question is, which of the images from the pool set would bring the most useful information to the model if annotated and added to the training set? The Train, Validation, and Test sets serve their usual roles. Train and Validation sets come from the first batch of images, while Pool and Test sets come from the second batch. The batches were scanned with different scanners. This shows how, in real life, we can have training and evaluation data coming from different distributions.

To obtain a baseline, I trained a single classification layer on top of a frozen ResNet18 feature extractor pre-trained on patches from TCGA-lung at 10x magnification [117] on the Train set using Cross-Entropy loss with 3 possible labels for the acinar pattern: "yes",

Train	Pool		**Accuracy**	**ROC AUC**	**Rank AUC**	**Pre (yes)**	**Re (yes)**
86		Train	0.65	0.82	0.752	0.333	0.2
86	10 x 10	Train + 10 Rand	0.75 ± 0.028	0.83 ± 0.026	0.82 ± 0.036	0.5 ± 0.191	0.39 ± 0.158
86	10	Train + 10 Rank	**0.8**	**0.864**	**0.874**	**0.667**	**0.6**
86	20 x 10	Train + 20 Rand	0.75 ± 0.04	0.83 ± 0.018	0.83 ± 0.042	0.59 ± 0.164	**0.46 ± 0.143**
86	20	Train + 20 Rank	**0.783**	**0.904**	**0.923**	**1.0**	0.2
86	30 x 10	Train + 30 Rand	0.76 ± 0.031	0.84 ± 0.017	0.85 ± 0.011	0.55 ± 0.097	0.57 ± 0.1
86	30	Train + 30 Rank	**0.817**	**0.924**	**0.954**	**0.833**	0.5
86	60	Train + Pool	**0.833**	0.865	0.894	0.636	**0.7**

Figure 3.11: Experiments were conducted by training the model on different training sets. For each $N \in \{10, 20, 30\}$, adding N ranked samples results in better models than adding N random samples. In both cases, the retrieved samples are annotated before being added to the training set. For random selection, 10 sets of N samples were taken from the Pool set, with mean $\pm$ one standard deviation reported for each metric.

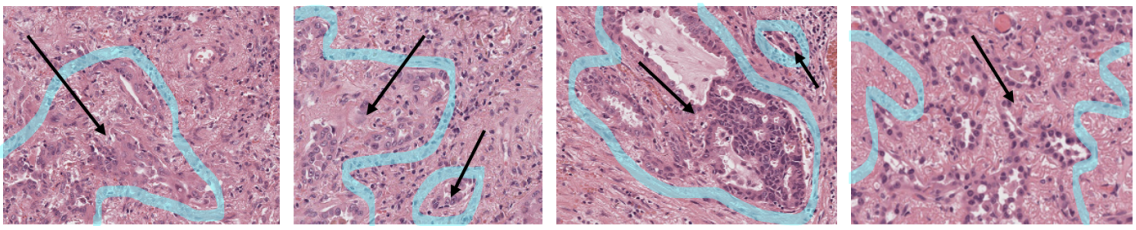

Figure 3.12: Regions containing acinar pattern from the top-10 ranked Pool set samples returned by our method. Solid arrows point at areas confirmed and delineated by the pathologist to contain acinar patterns, thus validating the results.

Set	Batch	Samples	Yes	No	Not Sure	Yes Proportion
Train	1	86	20	31	35	0.233
Validation	1	59	14	21	24	0.237
Test	2	60	10	46	4	0.167
10 Ranked Pool	2	10	4	5	1	0.4
Train + 10 Ranked Pool	1+2	96	24	36	36	0.25
20 Ranked Pool	2	20	6	13	1	0.3
Train + 20 Ranked Pool	1+2	106	26	44	36	0.245
30 Ranked Pool	2	30	8	21	1	0.267
Train + 30 Ranked Pool	1+2	116	28	52	36	0.241
Pool	2	60	10	46	4	0.167
Train + Pool	1+2	146	30	77	39	0.205

Table 3.6: Data distribution of regions with acinar pattern in different region sets.

"no", "not sure". When calculating Cross-Entropy loss, I used the same weights for all classes since the label distribution changes in different batches, and I wanted to be able to predict all of them well in the end. I report unweighted accuracy, ROC AUC weighted by the number of support samples with each label, as well as Ranking Curve AUC described in Section 3.3.1, precision, and recall for the "yes" label. I saved the weights of the models which showed the best ROC AUC on the validation set. Weighted ROC AUC was chosen because it is sensitive to class imbalance, gives a good understanding of the model performance, and is a popular choice in the literature [55, 117, 124, 190, 197].

Having the baseline model, I varied the training data by including different portions of the Pool set into it. Given that the "yes" label was under-represented in the first batch and my particular interest in learning to predict it better, I proposed to rank the samples by sorting them in decreasing order of predicted probabilities of the "yes" label. I assessed how

including 10, 20, and 30 ranked samples from the Pool affected the performance metrics. For comparison, I repeated the experiments with 10, 20, and 30 randomly chosen samples from the Pool. To account for randomness when choosing a subset of samples, I repeated the experiments 10 times for each subset size. Finally, I included all 60 Pool set samples to get the largest training data baseline (results in Figure 3.11).

Including more ($0 \rightarrow 10 \rightarrow 20 \rightarrow 30$) annotated ranked or annotated non-ranked samples from the pool set into the training set increased both the performance of the classifier (weighted ROC AUC) and the ability of the classifier to rank the regions with acinar pattern higher than the ones without it (Ranking Curve AUC). Furthermore, adding 10 ranked samples improved ROC AUC, Ranking Curve AUC, and Precision more than adding 10, 20, or even 30 random samples. Finally, including all 60 Pool samples improved accuracy and recall but resulted in lower weighted ROC AUC, Ranking AUC, and Precision. I believe that this happened because of optimising the parameters of the network using a non-weighted Cross-Entropy loss, which optimises the weights better for more prevalent classes. The proportion of the "yes" class in Train + Pool is 0.205, while it is 0.25. 0.245, 0.241, for Train + 10, +20, +30 Ranked images, respectively (see Table 3.6). This change in proportion resulted in improved unweighted accuracy but removed the precise focus from the "yes" class, which hurt ROC AUC (sensitive to class imbalance), Ranking AUC and Precision, which are all measured for the "yes" class.

The distribution of regions with acinar pattern presented in Table 3.6 shows that in all ranked pool sub-samples (10, 20, or 30), regions with a negative label were present. The classifier, trained only on the Train set, that we used to rank the images in the Pool set, already "knew" to give a high classification score to images with an acinar pattern present, meaning that it would not learn a lot from the addition of these images into the training set. However, giving high classification scores for images without an acinar pattern was a series of mistakes. I believe that it is the addition of these difficult-to-predict negative samples, also known as hard negative mining, that is proving the most influential for the

classifier improvement upon the addition of ranked images from the pool set.

Figure 3.12 shows samples from the top-10 ranked samples from the Pool set. The samples were ranked using predicted probabilities of the acinar pattern presence. These four regions were confirmed and delineated by the pathologist to contain the acinar pattern, thus adding further validation to my method.

3.4.3 Feature Space Investigation

The relative success of the TCGA pre-trained feature extractor on histopathology image patches at 10x magnification [117], compared to one pre-trained on natural images, motivated me to visualise patch features within the regions of interest in feature space. For visualisation, I extracted all 224×224 patches at 10x magnification. Patches intersecting diagnostically relevant ROIs were labelled as diagnostic; those intersecting benign ROIs were labelled as benign. All others were marked as "unknown". Diagnostic and benign ROIs cannot intersect by definition, as benign ROIs contain no cancerous tissue. Although such intersections did not occur in our data, they are theoretically possible, in which case the patch would be assigned the diagnostic label, consistent with the definition.

Figures 3.13, 3.15, and 3.14 show the first 2 UMAP [130] components of patch features. UMAP is a nonlinear dimensionality reduction technique that excels at visualising high-dimensional data. Compared to another standard dimensionality reduction technique, Principal Component Analysis (PCA), UMAP is preferred for high-dimensional (512 in our case) feature vectors because it better preserves both local and global data structures, reveals clusters more clearly, and handles complex, non-linear relationships, making visualisations more meaningful and insightful.

Figure 3.13 illustrates that in some slides, the feature extractor was able to achieve a successful separation of patches extracted from Diagnostic and Benign regions. This, however, was not true for all slides.

Figures 3.14 and 3.15 show feature vectors from the annotated diagnostic and benign

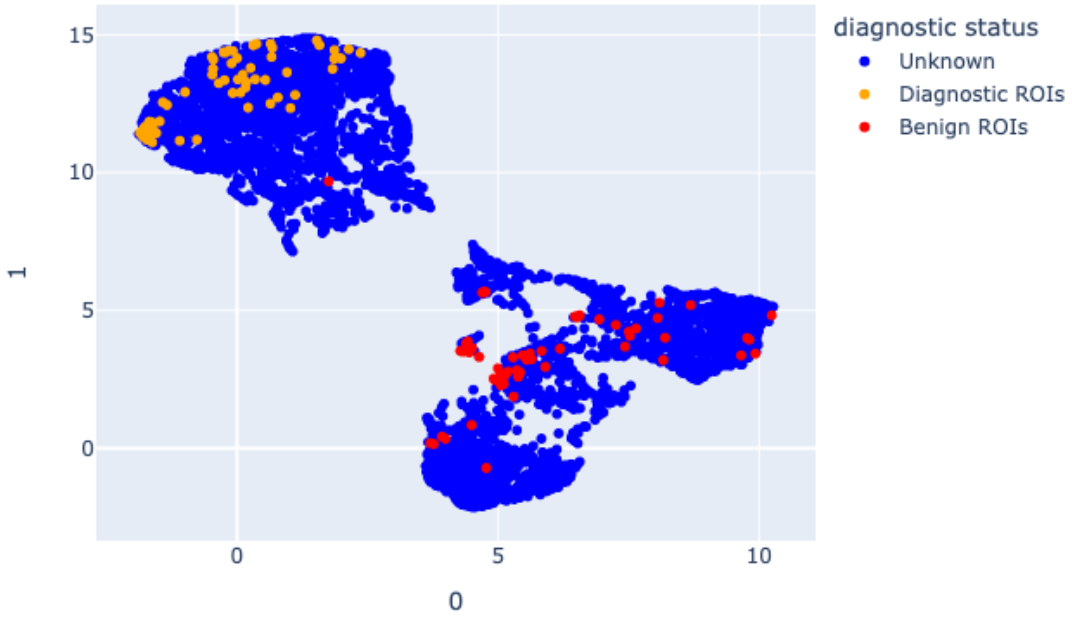

Figure 3.13: First 2 UMAP [130] components of patch features extracted with TCGA pre-trained 10x extractor [117]. Each patch corresponds to a coloured dot based on the ROI label it is coming from: diagnostic (orange), benign (red), or unmarked (blue).

patches. While there is some overlap between diagnostic tile features from different cancer subtypes, the feature extractor showed promising separation. In contrast, the feature extractor was not able to separate the patches marked as benign. This result is promising in a number of ways. The feature extractor trained only on Adenocarcinomas (LUAD) and Squamous Cell Carcinomas (LUSC) has learnt the lung cancer morphology enough to separate the diagnostic patches from three classes: LUAD, LUSC, Typical Carcinoids (TC). Regions marked as diagnostic provide the information needed for subtyping, while regions marked as benign provide good examples of subtype-agnostic tissue. The latter means that this annotation method can be used to explicitly supervise which regions to pay attention to and which regions to ignore when predicting the subtype of lung cancer.

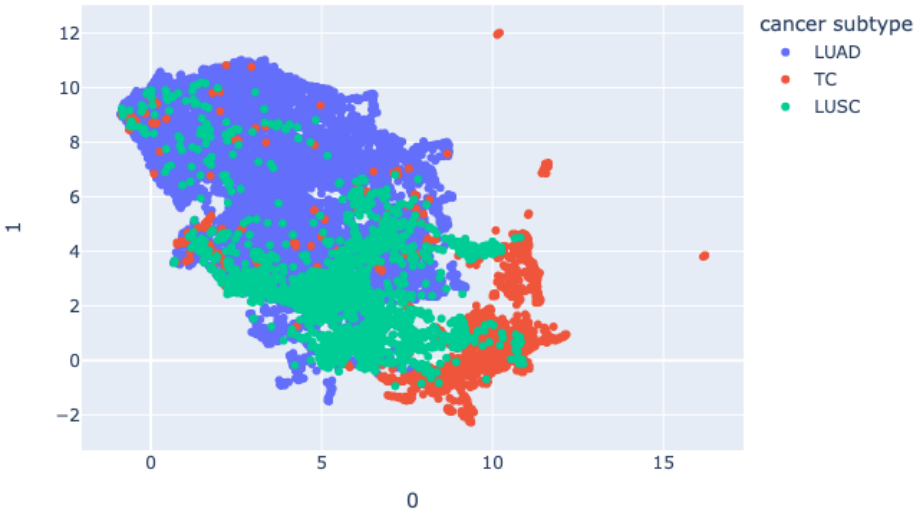

Figure 3.14: Features of diagnostic patches from the TCGA pre-trained 10x extractor [117] projected onto the first 2 components with UMAP [130]. Each tile corresponds to a dot, which is coloured by the lung cancer subtype present on the slide. Adenocarcinoma (LUAD), Squamous Cell Carcinoma (LUSC), Typical Carcinoid (TC).

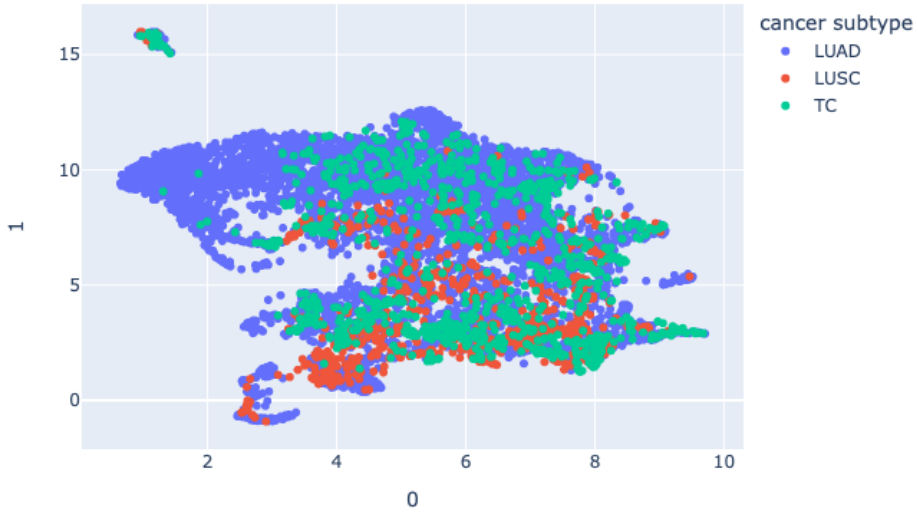

Figure 3.15: Features of benign patches from the TCGA pre-trained 10x extractor [117] projected onto the first 2 components with UMAP [130]. Each tile corresponds to a dot, which is coloured by the lung cancer subtype present on the slide. Adenocarcinoma (LUAD), Squamous Cell Carcinoma (LUSC), Typical Carcinoid (TC).

3.5 Results: Selecting Regions to Annotate

When we received the first region selections from the pathologists, I was excited by the possibility of automatically selecting diagnostic regions from slides. The idea was that I could systematically sweep through all patches on scanned WSIs to look for rare patterns. Systematic sweep is something the pathologist has no time to do. For example, some patches on the WSI can have the *keratinization* pattern, for which at the time there was only one sample, but if the pathologist does not decide to use these regions for the diagnosis, they would not be selected during the annotation process described in Section 3.2.1. If successful, this method would facilitate an N-fold increase for the rare patterns with very little additional effort from the pathologist. The downside of this method is that while improving it, pathologists might be asked to annotate clinically irrelevant regions, which would be a waste of the pathologists' time. The process of automatic selection of diagnostic regions was almost the same as described in Section 3.4. The only difference was that the potential pool of patches was not limited to the regions pre-selected by the pathologists but instead included all regions of the same physical size as the *query region* - a region annotated to have an under-represented pattern. The similarity of a candidate vector to the query vector was scored using the cosine similarity metric. Figure 3.6 (left) illustrates the one-shot retrieval process based on the similarity to the query sample. Enriching the dataset in this fashion poses a risk of only annotating new samples that are similar to the query samples, limiting the diversity of the dataset and preventing the models trained on this dataset from learning a full data distribution. However, if the feature extractor is weak or if the annotation process has just begun, the accidentally-retrieved samples with a negative label for the class of interest would benefit the learning by clarifying the decision boundary in the feature space between positive and negative labels.

At the time of this experiment, no feature extractors were pre-trained on pathology images in the public domain. However, [124] showed that one could use a truncated version of an ImageNet pre-trained ResNet50 to extract features from pathology images and

get state-of-the-art results on downstream tasks. During pre-processing, I resized the smaller dimension of the query region to 224 pixels and cropped the central 224×224 square, recording the resulting physical-to-pixel resolution. For candidate regions, I used a 224×224 pixel sliding window at the same physical-to-pixel resolution as the query region. Although preserving the resolution means that the feature extractor will be exposed to the same scale for both query and candidate regions, it also makes the method computationally expensive since, for each new query image, the process of feature extraction from the slide patches needs to be redone. Following the procedure described by Lu et al. [124], I normalised the images using ImageNet mean and standard deviation constants before passing the images into the feature extractor. I used a stride of $224/4 = 56$ to balance execution speed and match accuracy. Since the stride was smaller than the image size, I employed a procedure similar to non-max suppression and removed the shortlisted matches with $IoU \geq 0.25$ with any of the better matches.

I extracted the 20 best matches and asked the pathologists to confirm whether a keratinization pattern was present. Unfortunately, keratinization was not present in any of the candidate regions.

This negative result teaches several lessons. First, the feature extractor that helped Lu et al. [124] achieve state-of-the-art performance on the downstream task was not good enough for this task. Second, I should have tried to optimise the more time-consuming part of the annotation process as I did later on by prioritising regions to annotate (see Section 3.4). Finally, I should have considered the fact that annotating pre-selected ROIs in random order results in useful annotations, while annotating automatically selected regions that can present no clinical interest might waste already limited annotation time.

Many pathology-specific feature extractors have been released after my attempt (see Sections 2.3.3 on feature extraction from pathology images and 2.3.4 on pathology foundation models). My own experiments described in Sections 3.4.1 and 4.3 show the superior performance of pathology-specific pre-training compared to using features from models

pre-trained on natural images. Maybe using one of those models would have improved the results and enabled this step. However, none of these models were available at the time this experiment was done (autumn 2021). In 2024, Qiu et al. [148] showed that it is possible to efficiently mine regions from slides that are similar to query (prototype) regions from a visual-language database like ARCH [79] (curated pairs of medical images from pathology books and articles) or OpenPath [96] (pathology images with comments from Twitter) using PLIP [97] - a visual-language tile-level model pre-trained on OpenPath.

I hope that other researchers learn from my experience and consider the points above before trying to automate a task that is not the bottleneck of the process.

3.6 Results: Selecting Slides to Annotate

Models for extracting not only the patch but also the slide-level embeddings have been released in 2024. This enabled me to extend the idea of prioritising regions for annotation to enrich for under-represented morphological patterns into the idea of prioritising slides for annotation to enrich for under-represented cancer subtypes in scenarios when slide-level subtyping labels are unavailable. I used Prov-GigaPath [195] and PRISM [160] pre-trained slide aggregators available at the time. Prov-GigaPath represents purely vision pre-training, while PRISM exemplifies vision+language pre-training. I conducted the experiments on a small number of slides (the first 212 H&E slides from Oxford University Hospitals, see Section 3.2.3) to simulate the situation at the start of the annotation process when I would need this active enrichment the most. Hence, a zero-shot (using a caption) or one-shot (using a single query image) classification methods were the only options since training a supervised classifier for enrichment requires more data. The results of training linear classifiers on top of Prov-GigaPath and PRISM slide embeddings are discussed in Chapter 5. A pure vision slide feature extractor can be used for mining similar slides in exactly the same way as a patch feature extractor described in Section 3.4.1. A vision and language model can be used to generate slides' captions or compute

similarity scores with captions for different classes. Both models come with their respective tile-level feature extractors, which I used on all tissue patches extracted at 0.5 microns per pixel and normalised using ImageNet constants as instructed by the authors of PRISM [160] and Prov-GigaPath [195].

As described in Section 3.2.3 and shown in Tables 3.1 and 3.2, the digitised slides from Oxford University hospitals came in 3 batches of 20, 17, and 175 H&E images. The first batch was scanned using a scanner different from the last two. The third batch was scanned by a different person and at a higher magnification (40x instead of 20x) than the first two batches. Hence, when reporting the results, I report all combinations of batches 1, 2, and 3 to show how differences in slide processing can impact the model performance. The tissue of OUH slides came from lung adenocarcinoma (LUAD, 138 slides), squamous cell carcinoma (LUSC, 24 slides), typical carcinoid (TC, 41 slides), and other tumours (6 slides), while 4 slides presented only benign tissue. This data distribution suggests focusing on LUAD, LUSC, TC, and Other (benign slides and slides from other cancer types) categories.

3.6.1 Ranking Based on Similarity in Feature Space

I follow a similar pipeline as described in 3.4.1 to rank slides based on their similarity in feature space. However, I extended the evaluation strategy. First, I extracted slide embeddings for each slide and computed cosine similarities between all embedding pairs. After that, I considered each slide in turn as a query and evaluated how well the rest of the slides could be ranked based on the cosine similarity of their embeddings to the query embedding. I consider the ranking to be better if the slides with the same cancer subtypes as the query slide are ranked higher. For each query slide, I computed the area under the Ranking Curve (Ranking AUC) as described in Section 3.3.2 and the expected value of the Ranking AUC. Since each slide came from one of the four groups (LUAD, LUSC, TC, Other), I report the results per group in Tables 3.7, 3.8 and 3.9 to see how well I could mine slides with this particular subtype, no matter which slide I chose as the query.

OUH Batch	Model	LUAD	LUSC	TC	Other
1	PRISM	**0.833**		0.2	**1.0**
1	Prov-GigaPath	**0.833**		**0.6**	**1.0**
2	PRISM	**0.833**	0.143	**1.0**	
2	Prov-GigaPath	0.5	0.286	**1.0**	
3	PRISM	0.458	**0.688**	**0.781**	**0.857**
3	Prov-GigaPath	**0.508**	**0.875**	0.344	**0.714**
1 + 2	PRISM	**0.889**	**0.571**	**0.667**	**1.0**
1 + 2	Prov-GigaPath	**0.667**	**0.714**	0.444	**1.0**
1 + 3	PRISM	0.409	**0.688**	**0.811**	**0.9**
1 + 3	Prov-GigaPath	**0.727**	**0.938**	0.27	0.5
2 + 3	PRISM	0.484	**0.652**	**0.806**	**0.857**
2 + 3	Prov-GigaPath	**0.571**	**0.783**	0.333	**0.714**
1 + 2 + 3	PRISM	0.457	**0.565**	**0.829**	**0.9**
1 + 2 + 3	Prov-GigaPath	**0.79**	**0.826**	0.366	**0.6**

Table 3.7: Proportion of slides from each class, which, when used as queries, resulted in better Ranking AUC based on the feature similarities than the Ranking AUC of the expected case. Values larger than 0.5 are displayed in bold font to show that, for this scenario, it would be better to choose a slide at random and use the model ranking than to pick slides at random. LUAD: lung adenocarcinoma, LUSC: lung squamous cell carcinoma (no samples in OUH batch 1), TC: typical carcinoid of the lung (no samples in OUH batch 2), Other: benign and other under-represented subtypes.

Table 3.7 shows the proportion of query slides for each of the four labels that resulted in a higher Ranking AUC than the expected Ranking AUC from a random ordering. Most dataset-model configurations showed values larger than 0.5 (bold font in the table), meaning that you are more likely to select a query region that would result in a better-than-random ranking than a query region that would result in a worse-than-random ranking. Thus, it is advisable to choose a random query slide from the class of interest and use it as a query vector rather than order the slides randomly.

Tables 3.8 and 3.9 show the mean $\pm$ standard deviation Ranking AUC and the expected Ranking AUC of the random ordering for all four labels. The best-performing value for the label-dataset combination is highlighted in bold. Retrieval performance for typical carcinoid (TC) and Other groups was generally worse than for adenocarcinomas (LUAD) or squamous cell carcinomas (LUSC), with the PRISM model showing better performance than Prov-GigaPath. Having no access to the training sets for either of the models,

OUH Batch	Model	LUAD	E[LUAD]	LUSC	E[LUSC]
1	PRISM	**0.783 ± 0.103**	0.661	-	-
1	Prov-GigaPath	0.760 ± 0.129	0.661	-	-
2	PRISM	**0.667 ± 0.165**	0.542	0.486 ± 0.050	**0.562**
2	Prov-GigaPath	0.534 ± 0.100	0.542	0.496 ± 0.087	**0.562**
3	PRISM	**0.741 ± 0.077**	0.733	0.508 ± 0.063	0.503
3	Prov-GigaPath	0.737 ± 0.028	0.733	**0.659 ± 0.098**	0.503
1 + 2	PRISM	**0.679 ± 0.079**	0.608	**0.534 ± 0.042**	0.512
1 + 2	Prov-GigaPath	0.625 ± 0.065	0.608	0.530 ± 0.075	0.512
1 + 3	PRISM	0.734 ± 0.072	0.727	0.517 ± 0.048	0.503
1 + 3	Prov-GigaPath	**0.747 ± 0.028**	0.727	**0.688 ± 0.095**	0.503
2 + 3	PRISM	**0.726 ± 0.083**	0.714	0.505 ± 0.046	0.506
2 + 3	Prov-GigaPath	0.719 ± 0.024	0.714	**0.561 ± 0.082**	0.506
1 + 2 + 3	PRISM	0.721 ± 0.076	0.71	0.513 ± 0.037	0.505
1 + 2 + 3	Prov-GigaPath	**0.728 ± 0.024**	0.71	**0.584 ± 0.089**	0.505

Table 3.8: Ranking AUC retrieval performance of PRISM and Prov-GigaPath based on the feature similarities. LUAD: lung adenocarcinoma, LUSC: lung squamous cell carcinoma. E[X] is the expected ranking AUC if choosing samples at random to enrich for class X. OUH batch 1 did not have any slides with squamous cell carcinoma (LUSC).

I can only speculate about the reasons for the difference in performance. It is possible that training data for PRISM included more slides from these groups than the training data for Prov-GigaPath; however, it is also possible that the training data used for language grounding of PRISM put a lot of focus on TC and Other groups, partially compensating for their low prevalence in the dataset. For LUAD and LUSC slides, the retrieval performance is better, but I was unable to select a clear winner model, which is consistent with these two classes being more represented in public datasets [4, 10, 190].

Recall that batch 1 was scanned with a different scanner than batches 2 and 3. Examining the scores for batches (1, 3, 1 + 3) and (1, 2, 1 + 2) shows that the performance on a combination is comparable to the worse of the two performances, suggesting that both models can be used for extracting features from slides that come from two scanners without a retrieval performance drop compared to extracting features from slides scanned on a single scanner.

3.6.2 Ranking Based on Vision-Language Similarities

Using the PRISM model [160] enabled me to generate slide captions and compute similarities of different slides with a specified caption. The first capability can be used for zero-shot classification, while the second suits the ranking task better. Hence, I created a caption for each of the 3 classes: "lung adenocarcinoma" (LUAD), "squamous cell carcinoma" (LUSC), and "typical carcinoid" (TC) and computed a similarity score of each slide to each of the captions. I did not include the "Other" group due to the difficulty of defining a suitable caption. These similarities to the captions were used to rank images: the higher the similarity, the higher the ranking.

I also tried using multiple captions per label, using the maximum similarity of these captions with the slide. I used the same captions for LUAD and LUSC classes as described in the TCGA OncoTree used by Ding et al. [67] for their TITAN model: LUAD: "lung adenocarcinoma", "adenocarcinoma of the lung", "pulmonary adenocarcinoma", "peripheral lung adenocarcinoma", and "LUAD"; LUSC: "squamous cell carcinoma", "lung squamous cell carcinoma", "squamous carcinoma of the lung", "LUSC"; TC: "typical carci-

OUH Batch	Model	TC	E[TC]	Other	E[Other]
1	PRISM	0.340 ± 0.098	0.518	$\mathbf{0.806 \pm 0.196}$	0.503
1	Prov-GigaPath	$\mathbf{0.604 \pm 0.238}$	0.518	0.713 ± 0.116	0.503
2	PRISM	$\mathbf{0.911 \pm 0.024}$	0.512	-	-
2	Prov-GigaPath	0.689 ± 0.025	0.512	-	-
3	PRISM	$\mathbf{0.573 \pm 0.106}$	0.515	$\mathbf{0.587 \pm 0.100}$	0.5
3	Prov-GigaPath	0.479 ± 0.070	0.515	0.575 ± 0.121	0.5
1 + 2	PRISM	$\mathbf{0.636 \pm 0.118}$	0.522	$\mathbf{0.900 \pm 0.101}$	0.501
1 + 2	Prov-GigaPath	0.497 ± 0.113	0.522	0.781 ± 0.149	0.501
1 + 3	PRISM	$\mathbf{0.566 \pm 0.101}$	0.517	$\mathbf{0.582 \pm 0.094}$	0.501
1 + 3	Prov-GigaPath	0.463 ± 0.066	0.517	0.481 ± 0.065	0.501
2 + 3	PRISM	$\mathbf{0.604 \pm 0.124}$	0.516	$\mathbf{0.595 \pm 0.100}$	0.5
2 + 3	Prov-GigaPath	0.495 ± 0.066	0.516	0.581 ± 0.122	0.5
1 + 2 + 3	PRISM	$\mathbf{0.595 \pm 0.115}$	0.518	$\mathbf{0.592 \pm 0.099}$	0.501
1 + 2 + 3	Prov-GigaPath	0.477 ± 0.058	0.518	0.486 ± 0.066	0.501

Table 3.9: Ranking AUC retrieval performance of PRISM and Prov-GigaPath based on the feature similarities. TC: typical carcinoid of the lung, Other: benign and other underrepresented subtypes. E[X] is the expected ranking AUC if choosing samples at random to enrich for class X. OUH batch 2 did not have any slides with typical carcinoids (TC).

OUH Batch	LUAD	E[LUAD]	LUSC	E[LUSC]	TC	E[TC]
1	0.625	**0.674**	-	-	**0.63**	0.526
2	0.292	**0.555**	**0.717**	0.577	**0.63**	0.522
3	0.624	**0.734**	**0.654**	0.504	**0.541**	0.516
1 + 2	0.445	**0.615**	**0.771**	0.516	**0.555**	0.527
1 + 3	0.619	**0.729**	**0.656**	0.503	**0.53**	0.518
2 + 3	0.585	**0.715**	**0.673**	0.507	**0.553**	0.517
1 + 2 + 3	0.584	**0.711**	**0.678**	0.506	**0.541**	0.518

Table 3.10: Ranking AUC retrieval performance of PRISM based on the similarity of slide and caption embeddings. Captions used for classes: LUAD - "lung adenocarcinoma", LUSC - "squamous cell carcinoma", TC - "typical carcinoid". E[X] gives the expected ranking AUC if samples are chosen at random to enrich for class X. The bold font represents the winning strategy for a label-dataset combination. Example: when retrieving typical carcinoid slides from batch 1, it is better to rank images based on their similarity to a caption rather than using a random ordering.

noid", "typical carcinoid of the lung", "lung typical carcinoid". This approach produced similar but slightly inferior results, so I do not report them here.

Table 3.10 shows that ranking image-caption similarities led to better-than-random retrieval performance for squamous cell carcinoma and typical carcinoid slides, while performance for lung adenocarcinomas was consistently worse than random across all batches. This suggests that the model's vision encoder successfully captured features distinguishing adenocarcinoma presence, but the language encoder failed to reflect this in the captions. As a result, the model effectively ranked adenocarcinoma-negative slides higher than adenocarcinoma-positive ones—an inversion that can be corrected by reversing the PRISM ranking to recover useful performance.

Figure 3.16 visualises these patterns on the combined dataset (212 slides from OUH batches 1, 2, and 3). For squamous cell carcinoma (bottom left), ranking consistently outperformed random ordering. For typical carcinoid (bottom right), the ranking was generally better than random but less stable, occasionally dipping below the expected curve, particularly at the start and end of the ranked list. For adenocarcinoma (top left), the PRISM ranking curve remained consistently below the random baseline, while the reverse ranking (top right) outperformed it, confirming the hypothesis that reversing the

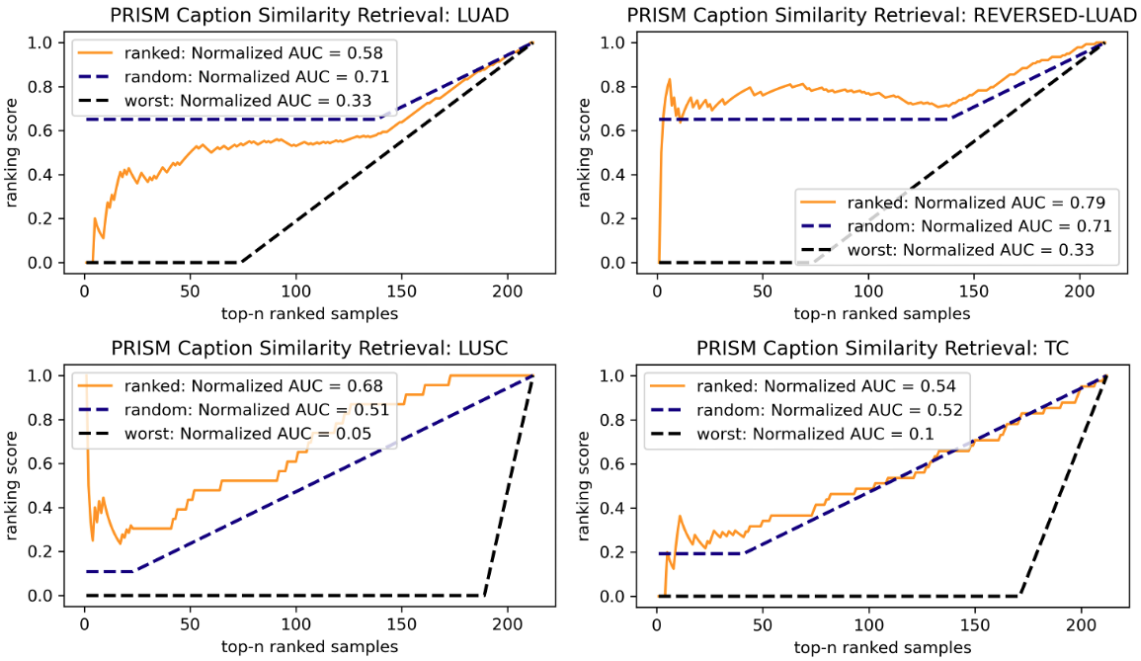

Figure 3.16: Ranking curves for the combined dataset (OUH batches 1, 2, and 3). Captions used for classes: LUAD - "lung adenocarcinoma", LUSC - "squamous cell carcinoma", TC - "typical carcinoid". For REVERSED-LUAD, the caption and the labels are kept the same as for LUAD, while the ranking is reversed.

ranking direction yields a practical retrieval strategy.

Overall, these results show that while the PRISM caption-similarity ranking is not uniformly reliable across classes, it can still be used effectively when its behaviour is well understood. When used for caption-based dataset enrichment, the PRISM model performs reliably for squamous cell carcinomas and shows potential for typical carcinoids. For adenocarcinomas, reversing the ranking provides a simple yet effective adjustment that enables enrichment beyond random selection.

3.7 Conclusions

Motivated by the ambition to link radiology and pathology imaging modalities for lung cancer patients, I developed a comprehensive annotation protocol for lung cancer histology images containing slide subtyping, diagnostic region selection, and region annotation stages. This protocol was used to annotate a new dataset I collected as part of the DART

Lung Health Project. The protocol was modified over time to accommodate the changes in the ongoing data collection that stretched from the beginning to the end of my DPhil.

Due to the nature of ongoing data collection, the dataset was unbalanced at all times. Some lung cancer subtypes were present on many slides, while others were present on very few. The pathologists did not have time to annotate all slides to the extent we originally planned. To use the expert annotation time more efficiently, I proposed one-shot retrieval and supervised methods for (1) choosing which of the pre-selected regions to annotate, (2) selecting diagnostic regions from slides, and (3) selecting which of the slides should be annotated next. To be able to evaluate these methods in a limited annotation time setting on the new dataset described in this chapter, I proposed a new metric, Ranking AUC, which was praised by the MICCAI 2022 CaPTion workshop reviewers as "an excellent attempt to formalise some of their points using a mathematical formula".

1) The proposed method effectively prioritises annotating regions extracted from whole-slide images likely to contain under-represented patterns in one-shot retrieval and supervised settings. In the one-shot setting, it enabled efficient retrieval of regions. In a supervised setting, adding ranked images improved the performance of pattern classifiers compared to adding random images, which can possibly be attributed to hard negative mining. I conclude that even with little supervision, the dataset can be enriched for a pattern of interest. This work was presented at the MICCAI 2022 CaPTion workshop and at the Digital Pathology showcase organised by Roche at the Royal College of Pathologists.

2) Although my attempt at selecting diagnostic regions similar to a query region containing under-represented patterns was unsuccessful in 2021 using the available feature extractor models, it might bring promising results now. A work from 2024 by Qiu et al. [148] showed that modern visual-language extraction models can facilitate efficient mining of regions with specific morphologies.

3) Slide-level foundation models (PRISM [160] and Prov-GigaPath [195]) developed in 2024 proved to be effective for ranking new slides based on a feature similarity to a query

slide with a known lung cancer subtype. Using the vision-language capabilities of PRISM also proved effective for prioritising squamous cell carcinoma, and moderately effective but less reliable for typical carcinoid slides. For adenocarcinoma slides, the reversed caption similarity ranks provided better-than-random performance, correcting the failure of the original ranking direction. More slide-level foundation models released in 2024 (TANGLE [104] and TITAN [67]) could be added to the evaluation as part of future work. However, the focus of this part of the chapter was to show the feasibility of the proposed method and not to benchmark all available slide-level models.

Having shown that using pathology-specific feature extraction models can enable better prioritisation of pre-selected regions, I explore the question of selecting a promising feature extractor to use on custom datasets in Chapter 4.

The new dataset and slide-level embedding models described in this chapter are used in Chapter 5 for developing a lung cancer subtyping model.

Chapter 4

Evaluating, Selecting, and Using Pathology Foundation Models

Contents

Access to publicly available digital histopathology datasets has driven significant progress in developing deep learning-based frameworks for histopathology. A notable advancement in recent years is the emergence of foundation models, pre-trained on massive histopathology datasets in a self-supervised manner. While public whole slide image datasets initially played a critical role, academic and industry research groups have increasingly turned to proprietary datasets to train their foundation models, relegating public datasets to evaluation purposes.

In recent years, datasets composed of image patches have been predominantly used for evaluation. The LC25000 dataset, consisting of tissue image tiles extracted from lung and colon samples, has become a widely used benchmark. In its released form, tissue tiles were randomly augmented and mixed, leading to near-perfect accuracy scores in many studies, often due to data leakage caused by augmented versions of the same tile being split across training and test sets. To address this issue, I developed a semi-automatic pipeline to clean the LC25000 dataset. By clustering and separating all augmented versions of the same tiles using recently proposed histopathology foundation models, followed by manual correction, I created a clean version of the dataset. This allowed me to evaluate the quality of features extracted by foundation models through clustering tasks as a benchmark. Subsequently, I trained simple classifiers on these features, concluding that classification tasks on LC25000 have been solved by the latest foundation models, none of which had the LC25000 dataset used for training, according to the authors.

Interestingly, models that performed well at clustering augmented patches also excelled in the classification task. This observation motivated me to explore clustering augmented patches extracted from whole slide images as a pretext task to identify a promising foundation model for feature extraction rather than exhaustively computing all patch features with every available foundation model for downstream whole-slide classification tasks. This approach is conceptually related to contrastive and self-distillation-based methods like SimCLR [49] and DINO(v2) [41, 141], which aim to learn meaningful representa-

tions by encouraging consistency across different views of the same image.

Contributions

LC25000-clean: Minimal Histopathology Benchmark

1. Publicly released our semi-automatic annotation pipeline along with the LC25000-clean dataset to facilitate appropriate utilization of this dataset, reducing the risk of overestimating models' performance;
2. Profiled combinations of feature extraction and clustering methods for finding duplicates of the same image generated by basic image transformations;
3. Proposed the clustering task as a minimal-setup benchmark to evaluate the quality of tissue image features learned by histopathology foundation models. Clustering labels, annotation pipeline, and evaluation code: `https://github.com/GeorgeBatch/LC25000-clean`

Selecting a Candidate Model for your WSI data

1. Proposed to use clustering augmented patches as a pretext task to choose a promising foundation model for new whole slide datasets: `https://github.com/GeorgeBatch/histFM-clustering-benchmark`
2. Evaluated the ranking agreement between foundation model performance on different pretext tasks and downstream classification performance.
3. Added support for non-tiled ".tif" files produced by Roche Ventana scanners and UNI [46], Prov-GigaPath [195], and H-Optimus-0 [158] foundation models to Tiatoolbox library [145]: `https://github.com/TissueImageAnalytics/tiatoolbox/releases/tag/v1.6.0`

This chapter includes "Evaluating Foundation Models for Few-shot Tissue Clustering: an Application to LC25000 Augmented Dataset Cleaning" [34] paper, which received the Best Paper Award at the Data Engineering for Medical Imaging (DEMI) workshop of the Medical Image Computing and Computer Assisted Intervention (MICCAI) 2024 conference.

4.1 Introduction

Public digital histopathology datasets, consisting of digitized tissue specimens (some popular datasets in alphabetical order: BACH [21], CAMELYON [122], DHMC [190], PAIP [109], PANDA [38], TCGA [4], TCIA [53], etc.) have facilitated advancements in computational pathology, which is vital for developing and validating deep learning-based frameworks such as tumour detection, cancer subtyping, and therapy response prediction [182]. The availability of these public datasets enabled the creation of the early pathology-specific foundation models: TransPath [186] with its follow-up CTransPath [187] trained on TCGA in a self-supervised manner, and ResNet18 models pre-trained on TCGA lung and CAMELYON-16 datasets [117]. Since then, more academic and industrial research groups have chosen to collect proprietary datasets for pre-training (UNI [46], Prov-GigaPath [194], Virchow [185], H-Optimus-0 [158], and others), while some groups continued developing on public datasets (REMEDIS [24], Phikon v1 and v2 [73], PathDINO [19], and others). This shift was motivated by the need to comply with stringent medical-data regulations, protect intellectual property associated with proprietary datasets, and address concerns over *data leakage*, where exposing evaluation data to a model during training, even in a self-supervised context, can result in artificially inflated metrics that fail to reflect the model's true capabilities. Tile-level datasets, like the NCT-CRC-HE-100K (Kather-100k) dataset of colorectal cancer patches [108], have been used to evaluate these foundation models.

4.1.1 LC25000 Dataset

While public datasets enable significant progress, caution is necessary to avoid common pitfalls that may lead to data leakage and invalid model evaluation.

The LC25000 dataset [35], for instance, is a widely-used benchmark source, cited over 440 times[1]. It comprises 25,000 tiles extracted from lung and colon histopathology

[1] https://scholar.google.co.uk/scholar?cluster=14706626049620623468

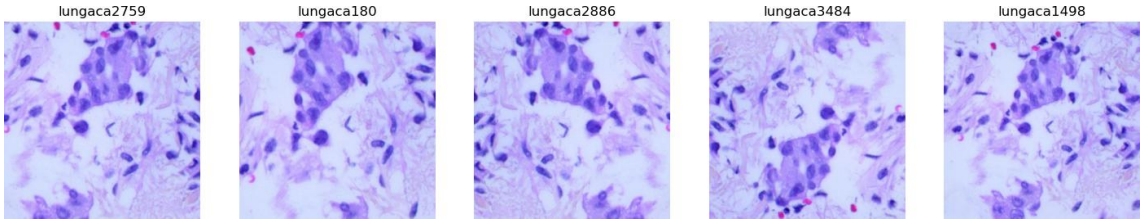

Figure 4.1: Examples of augmented images from the same origin tile with their names in the released dataset showing on the top. The images were shuffled and indexed randomly.

images, divided into five classes with 5,000 images per class: lung adenocarcinoma (lung_aca), lung squamous cell carcinoma (lung_scc), normal lung (lung_n), colon adenocarcinoma (colon_aca), and normal colon (colon_n). This dataset is mainly used in developing tile-level classifiers for cancer tissue classification, with various studies reporting accuracy scores of 95% and above [127, 129, 132, 169, 174, 199].

However, it is crucial to be aware of issues such as *type-1* data leakage, where the augmented images of the same tissue tile are split into the training and test set, leading to over-estimation of model performance attributing to simple shortcuts that associate these duplicates. Similar concerns apply to datasets like TCGA, where multiple slides from the same patient may inadvertently be included in both training and testing sets (*type-2* data leakage).

According to the original LC25000 publication [35], the authors collected 250 original images for each of the five classes mentioned above, which I will refer to as **prototypes**. They expanded the dataset to 25,000 images through a series of random image transforms, including random rotations and flips. All images were centre-cropped and resized to 768x768 pixels. Given that 5,000 images in each class are derived from 250 prototypes, each prototype generates around 20 augmented images. We refer to augmented images of the same tissue tile (a prototype) as **semantic duplicates**. Suppose these 5,000 images are randomly split into training, validation, and test sets using an 80/20 ratio. Then there is a $\sim 99\%$ chance $(1 - p(\text{all 20 in test}) - p(\text{all 20 in train}) = 1 - 0.2^{20} - 0.8^{20})$ semantic duplicates of the same prototype appearing in both the training and test sets. Consequently, it can be expected that *type-1* data leakage would occur for almost all prototypes when

models are both trained and evaluated on LC25000.

Acknowledging the contributions and widespread use of LC25000, I aimed to enhance the quality of performance reports when researchers utilize this dataset. Many studies report near-perfect accuracy scores (often 99.9%), primarily due to data leakage. This has, unfortunately, led to an unnecessary allocation of research resources and reviewing efforts. Therefore, I intended to support the research community in making more accurate and reliable use of the LC25000 dataset.

Inspired by the recent success of automatic data cleaning of the Quilt-1M histopathology dataset [23, 101], I developed a semi-automatic pipeline to clean the LC25000 dataset. In my approach, I first cluster the images into groups of images originating from the same prototype and then perform a semi-automatic check and correction, creating LC25000-clean dataset where all semantic duplicates belonging to the same prototypes are grouped together.

4.1.2 Contributions

The contributions can be split into two groups. The first group is limited to the LC25000 dataset. First, I release our semi-automatic annotation pipeline along with the LC25000-clean dataset to facilitate appropriate utilisation of the LC25000 dataset. Second, I assess various combinations of feature extraction and clustering methods for clustering semantic duplicates. Finally, I propose using the clustering task as a minimal-setup benchmark to evaluate emerging histopathology foundation models. This task can be regarded as a quality lower bound of the tissue image features extracted by these models.

The second group extends the ideas tested on the LC25000 dataset to 5 whole-slide image datasets. First, I propose to sample and augment patches from whole-slide images to create data for the clustering task. Second, I assess how effectively the augmented patches can be separated into groups using the raw features from 5 pathology feature extractors. Finally, I perform downstream classification tasks and analyse if the clustering

performance can inform the choice of the feature extractor for a new downstream task.

4.2 LC25000-clean: Semi-automatic Dataset Cleaning

The first goal was to obtain a clean version of LC25000, where semantic duplicates are grouped according to their prototype memberships. When I reached out[2], the authors replied that they did not have permission to share original data.

Feature extraction. The original 768x768 image tiles are resized into 224x224, and image features are computed using pre-trained image models, such as the UNI [47] model. The resulting features are used for clustering to produce an initial cleaning result. I also evaluated the effects of different image normalization options, feature extractors, and dimensionality reduction methods.

Image Pre-processing. I explored three different image RGB normalization options while always resizing the 768x768 images to 224x224: (1) resize-only: no normalization applied after resizing; (2) normalizing with ImageNet constants as suggested by the authors of UNI [47], Prov-GigaPath [195], and Phikon[73] models; (3) normalizing using mean and standard deviation constants computed separately for each of the 5 classes of LC25000.

Feature Extraction. I extracted features with models pre-trained on natural images and histopathology-specific models. Natural-image models included ResNet18 [86] and Truncated ResNet50 as used in the CLAM pipeline [124], as well as the small and base versions of ViT pre-trained with DINOv2 [141] strategy. Pathology-specific models included ResNet18 pre-trained with SimCLR [49] on the lung portion of TCGA [117], UNI [47], Prov-GigaPath [195], and Phikon [73].

Dimensionality Reduction. The feature extractors described above output feature vectors in different sizes: 512, 768, 1024, 1536. However, the information encoded might be

[2] https://github.com/tampapath/lung_colon_image_set/issues/5

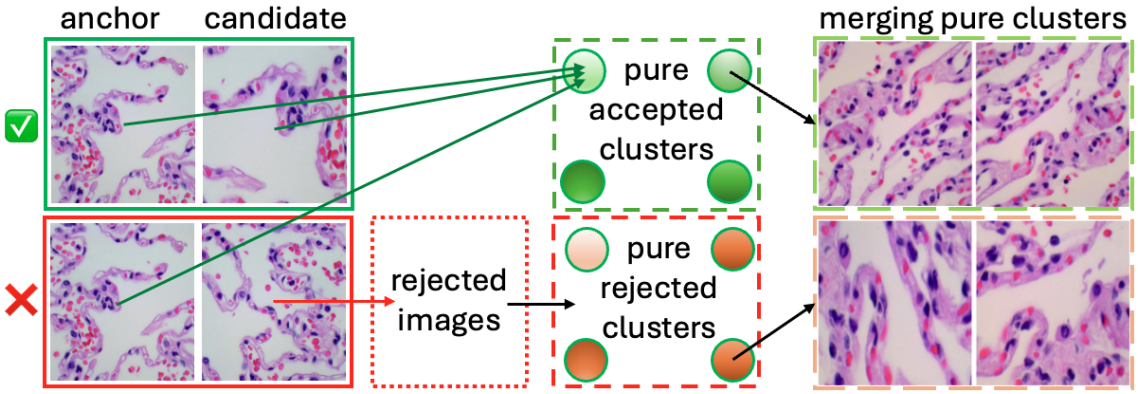

Figure 4.2: Manual Annotation Framework Schema. Top-left: accepted positive pair. Bottom-left: rejected negative pair. Rejected image is added to the pool of rejected images. After the initial stage is complete, all rejected images are clustered and the clusters are purified. Finally, pure accepted and rejected clusters are merged if needed.

excessive when clustering augmented images. Hence, I tested PCA and UMAP for reducing the dimension of extracted features before running the clustering algorithms: PCA with 0.9, 0.95, and 0.99 proportion of variance explained and UMAP [130] with 2, 8, 32 components (other parameters were fixed).

Clustering and manual cleaning. The feature vectors corresponding to the image tiles were fed into K-Means with 250 clusters per class, producing an initial clustering. The manual stage contained three components: (1) manually accepting or rejecting cluster assignments by iterating through the initial clustering result, (2) automatically clustering the rejected images and manually purifying rejected clusters, and (3) manually merging pure accepted and rejected clusters.

Manual Accepting and Rejecting of Cluster Assignments. First, I computed the centroid embedding of each cluster c and selected the image closest to the centroid as a reference I_r^c. Then, for each cluster c I displayed a pair $(I_r^c, I_j^c) \forall j \neq r$ and iterated over all other images I_j^c that have been automatically assigned to this cluster. The annotator was then prompted to choose whether an image comes from the same origin (see Figure 4.2, left). If so, image I_j^c is confirmed to belong to cluster c. Otherwise, I_j^c is added to the pool of rejected images. After reviewing all 250 clusters in the manner described above, I ended

up with accepted and rejected sets. The accepted set contained 250 pure clusters. The rejected set contained the images incorrectly assigned by the clustering algorithm (2% of image tiles for lung tissue, 3% for colon tissue).

Clustering Rejected Images and Purifying Rejected Clusters. I ran the clustering algorithm on the rejected images. To maximise the chance of obtaining pure clusters, into clusters of an average size of 10 (twice smaller than the average size of the original clusters of 20). Then, an annotator looks through and manually splits impure clusters into pure clusters until only pure clusters remain.

Merging Pure Accepted and Pure Rejected Clusters. Given that the clusters in the accepted and rejected pools are pure, the next step is to merge over-grained clusters (examples of clusters to be merged are shown in Figure 4.2, right). I computed pairwise distances between the clusters via single linkage, i.e., assigning the distance between clusters $\mathbf{A}$ and $\mathbf{B}$ with the smallest distance between any two pairs of feature vectors $\mathbf{a_i} \in \mathbf{A}$ and $\mathbf{b_j} \in \mathbf{B}$, and checked the closest pairs of clusters to decide if a pair of clusters should merge.

Judging by the visual purity assessment of the initial clusters, I settled on my final pipeline, which consists of feeding resized images directly into the UNI feature extractor, reducing the output features using UMAP with 8 components, and running K-Means on the reduced features.

4.3 Experiments and Results: LC25000

After obtaining LC25000-clean, where all images were grouped and assigned prototype labels, I evaluated the quality of features extracted using different histopathology foundation models and baseline feature extractors pre-trained on natural images by clustering the original LC25000 dataset and comparing the clustering results against LC25000-clean. Furthermore, I evaluated the classification performance of KNN and Linear classifiers on

top of the raw features extracted by a subset of the foundation models on LC25000 and LC25000-clean datasets. I observed that (1) pathology foundation models performed better than feature extractors pre-trained on natural images on both tasks, and (2) there was a direct correspondence between clustering and classification ranks of model performances, as shown in Figures 4.3, 4.4, 4.5 and Table 4.1.

4.3.1 LC25000-clean: Evaluation of Augmented Patch Similarities

I profiled the quality of image features learned by recent histopathology foundation models clustering LC25000 and tested the clustering results against LC25000-clean. I also tested the effects of different image normalization methods. In a perfect scenario, LC25000 can be automatically (without manual correction) clustered into LC25000-clean, where each cluster contains all semantic duplicates generated by random image transformation from the same prototype.

For the clustering assignment A_p, I computed a connectivity matrix between all pairs of images where any two images belonging to the same cluster are connected. I then evaluated this assignment against the ground truth assignment A_m. Binary connectivity (Accuracy, Precision, Recall, F1-score, Specificity, Balanced Accuracy) and clustering (Fowlkes–Mallows index [75], Adjusted Rand Index [98, 150], Normalized Mutual Information score, Homogeneity, Completeness, V-Measure [154]) are used to evaluate the clustering performance. I further incorporated precision@1 and precision@5 for evaluation, as they are agnostic to the clustering algorithm (e.g., precision@1 only checks the fraction of samples whose nearest neighbour's label is the same as the sample itself). Fowlkes–Mallows Index (**FMI**) is used as the main metric for assessing clustering performance, as it is a balanced metric and reflects the global clustering quality. To compute FMI, the best alignment of the two clustering assignments is found using the Hungarian Algorithm [113]. Then, the numbers of true positives (pairs of points that are in the same cluster in both A_m and A_p), false positives (pairs of points that are in the same cluster in A_p, but in different clusters in A_m), and false negatives (pairs of points that are in different

clusters in A_p, but in the same cluster in A_m) are calculated. Finally, FMI is calculated as $FMI = \sqrt{(TP/(TP+FP)) \times (TP/(TP+FN))} = \sqrt{Precision \times Recall}$.

The raw features extracted by each model were clustered using the K-Means algorithm. I noticed that the results varied for the same feature extractor depending on the image normalization method, with FM-index reported in Figure 4.4. Figure 4.3 shows the clustering performance under different metrics, with each feature extractor's best-performing image normalization technique (by FM-index).

UNI and Prov-GigaPath outperformed other models by large margins. Interestingly, the best image normalization method for UNI was no RGB normalization, whereas for Prov-GigaPath, it was to normalize the input using the corresponding dataset's statistics such that the dataset is centred and has a unit variance (i.e., the mean and variance of the LC25000 lung dataset). Note that both models were trained with inputs normalized using ImageNet statistics [47, 195].

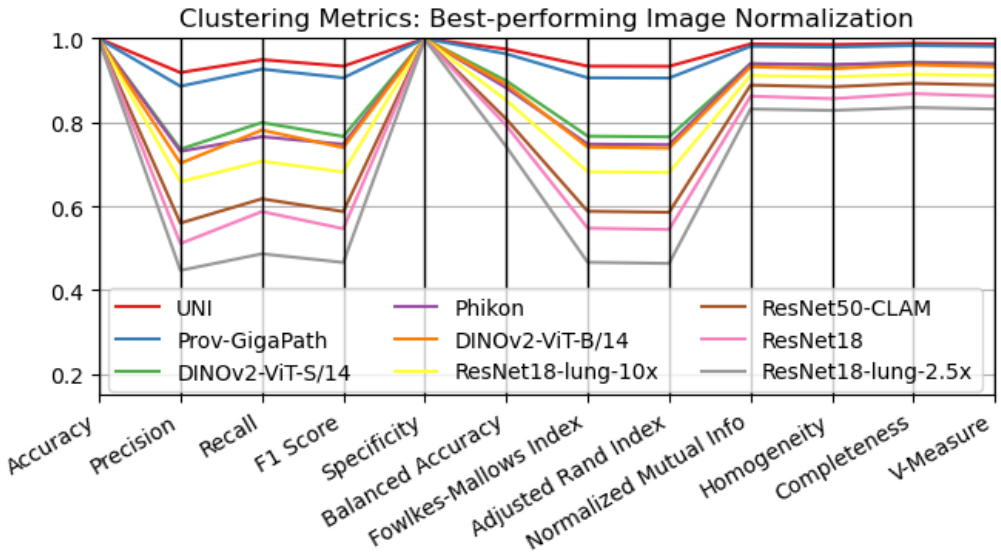

Figure 4.3: Clustering performance: all metrics. Outputs from feature extractors are passed directly into K-Means clustering. For each feature extractor, the results correspond to using the best normalization-extractor combination (by FM-index - see Figure 4.4). Note, UNI, Prov-GigaPath, Phikon, ResNet18-lung -10x and -2.5x were pre-trained on pathology images, while DINOv2-ViT -S/14 and -B/14, ResNet50-CLAM, and ResNet18 were pre-trained on natural images.

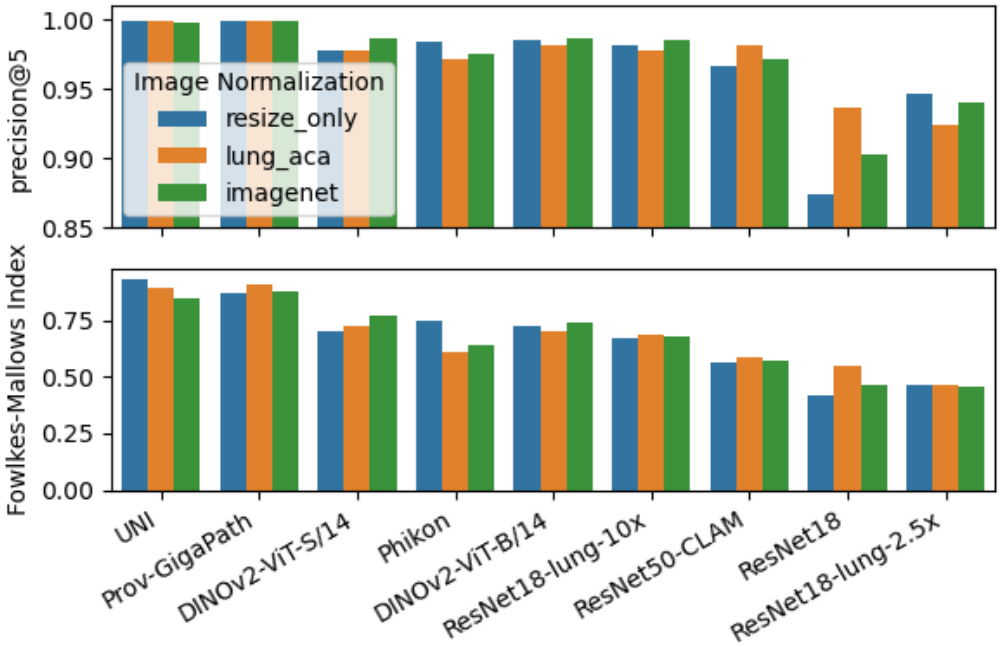

Figure 4.4: **Precision@5.** Outputs from feature extractors are ranked by how close they are in Euclidean distance. **Fawlkes-Mallows Index.** Raw outputs from feature extractors are passed into K-Means clustering (Euclidean distance). **Image Normalization methods.** *resize_only*: using raw RGB values (0 to 1). *lung_aca*: using the statistics computed from the LC25000 dataset (mean and variance) to centre the inputs and make a unit variance. *imagenet*: using the statistics of the ImageNet dataset.

4.3.2 LC25000-clean: Classification with KNN-1 and Linear Probing

LC25000 was originally proposed for two classification tasks: three-class lung tissue classification (adenocarcinoma, squamous cell carcinoma, normal lung tissue) and binary colon tissue classification (adenocarcinoma vs normal colon tissue). To evaluate the effect of *type-1* data leakage on the reporting performance of the classification model on the original LC25000 dataset, I trained classifiers using KNN and Linear probing based on raw image features extracted by the foundation models.

Randomly splitting the original LC25000 dataset images into train and test sets results in both sets containing semantic duplicates from the same prototypes (expected contamination of 99% with an 80/20 train/test split - see Section 4.1.1). In this circumstance, KNN with 1 nearest neighbour can achieve an almost perfect accuracy score, because, for each image, the nearest neighbour is always one of its semantic duplicates (thus, having the

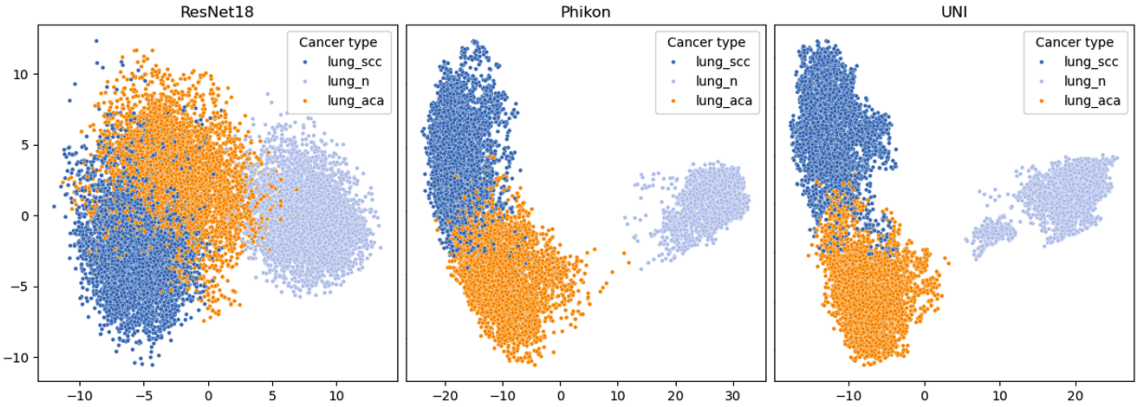

Figure 4.5: First 2 PCA projections of raw feature outputs of ResNet18, Phikon, and UNI feature extractors (images were normalized using ImageNet normalization constants).

same class label). For linear probing, I trained a single linear layer with softmax activation for the multi-class lung tissue classification and sigmoid activation for the binary colon tissue classification. Similar to the case of KNN, the linear layer should be able to overfit the training data, thereby overfitting the test data provided that most samples in the test set will have semantic duplicates in the training set. Therefore, I overfitted by training the model until the training loss converged (no validation).

In contrast, on the clean dataset, it is harder for the KNN classifier with 1 nearest neighbour to achieve perfect accuracy because no image in the test set is a semantic duplicate of any images in the training set. For the linear classifier, overfitting the training set is expected to result in a performance drop.

To clearly delineate how much of the previously reported performance is attributed to the data leakage, I added further stress tests by decreasing the ratio of training samples, starting with a popular 80/20 train/test split and decreasing the ratio to 5/95, where only 5% of samples were used for training.

Compared to ResNet18 pre-trained on natural images, lung tissue classes were better separated in feature space by Phikon and UNI, ViT-based models pre-trained with self-supervised methods (iBOT [204] and DINOv2 [141], respectively) on large pathology datasets, confirming the advantage of pathology-specific foundation models (Figure 4.5).

Split	Features	CLS	Lung		Colon	
			Original	**Clean**	**Original**	**Clean**
80% - 20%	ResNet18	KNN	0.998 ± 0.001	0.917 ± 0.007	0.999 ± 0.000	0.962 ± 0.007
		Linear	0.970 ± 0.003	0.943 ± 0.007	0.998 ± 0.001	0.993 ± 0.003
	Phikon	KNN	$\mathbf{1.0 \pm 0.0}$	0.993 ± 0.003	$\mathbf{1.0 \pm 0.0}$	$\mathbf{1.0 \pm 0.0}$
		Linear	$\mathbf{1.0 \pm 0.0}$	0.987 ± 0.009	$\mathbf{1.0 \pm 0.0}$	$\mathbf{1.0 \pm 0.0}$
	UNI	KNN	$\mathbf{1.0 \pm 0.0}$	0.994 ± 0.005	$\mathbf{1.0 \pm 0.0}$	$\mathbf{1.0 \pm 0.0}$
		Linear	$\mathbf{1.0 \pm 0.0}$	0.989 ± 0.01	$\mathbf{1.0 \pm 0.0}$	$\mathbf{1.0 \pm 0.0}$
20% - 80%	ResNet18	KNN	0.981 ± 0.002	0.892 ± 0.009	0.992 ± 0.001	0.935 ± 0.013
		Linear	0.961 ± 0.002	0.923 ± 0.007	0.995 ± 0.001	0.981 ± 0.002
	Phikon	KNN	$\mathbf{1.0 \pm 0.0}$	0.972 ± 0.006	$\mathbf{1.0 \pm 0.0}$	0.999 ± 0.001
		Linear	0.998 ± 0.000	0.977 ± 0.004	$\mathbf{1.0 \pm 0.0}$	0.999 ± 0.001
	UNI	KNN	$\mathbf{1.0 \pm 0.0}$	0.978 ± 0.005	$\mathbf{1.0 \pm 0.0}$	$\mathbf{1.0 \pm 0.0}$
		Linear	0.999 ± 0.000	$\mathbf{0.979 \pm 0.003}$	$\mathbf{1.0 \pm 0.0}$	0.999 ± 0.001
5% - 95%	ResNet18	KNN	0.940 ± 0.005	0.862 ± 0.020	0.970 ± 0.002	0.872 ± 0.038
		Linear	0.945 ± 0.002	0.894 ± 0.019	0.984 ± 0.002	0.942 ± 0.020
	Phikon	KNN	0.994 ± 0.001	0.948 ± 0.013	$\mathbf{1.0 \pm 0.0}$	0.994 ± 0.004
		Linear	0.993 ± 0.002	0.960 ± 0.005	$\mathbf{1.0 \pm 0.0}$	0.992 ± 0.005
	UNI	KNN	$\mathbf{0.996 \pm 0.001}$	$\mathbf{0.962 \pm 0.008}$	$\mathbf{1.0 \pm 0.0}$	$\mathbf{0.997 \pm 0.002}$
		Linear	0.994 ± 0.002	0.960 ± 0.014	$\mathbf{1.0 \pm 0.0}$	0.990 ± 0.008

Table 4.1: Classification accuracy (mean $\pm$ st.d.) computed on 10 random splits of ResNet18, Phikon, and UNI extractors with ImageNet image normalization.

Table 4.1 shows the classification accuracy achieved by KNN (1 nearest neighbour) and linear classifiers on the features computed from ImageNet pre-trained ResNet18, Phikon, and UNI feature extractors on original and cleaned versions of the LC25000 dataset (images were normalized using ImageNet constants, same for all models to facilitate the comparison). These 3 models are chosen based on the clustering performance of the lung adenocarcinoma class, and on that, they have different representation strengths: ResNet18 (weak), Phikon (medium), and UNI (strong), as indicated in Figure 4.4. I varied the train/test split proportions to show how easy/difficult it is to achieve near-perfect prediction accuracy on the original and cleaned versions of the dataset. The reported values were computed from 10 random draws for each split proportion, with the mean and standard deviation of the accuracy values.

I observed that the classification accuracy ordering of the feature extractors is consistent with the clustering performance: UNI performed best, Phikon followed, while ResNet18 pre-trained on ImageNet showed the worst results. For the original dataset, the models all

could achieve very high accuracy; even with only 5% of the images for training, a simple KNN classifier could still achieve a near-perfect performance (>99.5% accuracy), simply due to overfitting and data leakage. As expected, the classification accuracy suffered a statistically significant drop on the cleaned dataset compared to the original dataset, which is subjected to *type-1* data leakage, highlighting the importance of appropriate treatment of the dataset for a robust train/test split that reflects the true power and generalizability of the evaluated models. Notably, both Phikon and UNI still achieved strong performance ($\geq$95% accuracy) on the cleaned dataset with just 5% of the training data, suggesting that the LC25000-clean classification task is solved at the feature level by the foundation models which separate the classes in feature space as shown in Figure 4.5.

4.4 Experiments and Results: Whole Slide Datasets

In Section 4.3, I observed a correlation between the ability to cluster augmented patches and classification performance on the LC25000-clean dataset. This observation sparked a hypothesis that the same relationship might also hold between the ability to cluster augmented sample patches extracted from whole-slide images and downstream task performance on these whole-slide images. This hypothesis was received with excitement during my presentation of "Evaluating Foundation Models for Few-shot Tissue Clustering: an Application to LC25000 Augmented Dataset Cleaning" [34] at the MICCAI 2024 DEMI workshop. This raised an exciting possibility: if the hypothesis held, one could identify a promising foundation model without computing features for all patches in a whole-slide image dataset, significantly reducing computational cost. However, the experiments presented in this section suggest that clustering a sample of augmented patches is not a reliable proxy for selecting the best-performing foundation model on a new dataset. Whether this limitation stems from the specific pretext task or reflects a more fundamental difficulty in predicting downstream performance rankings remains an open question for future research.

4.4.1 WSI Datasets

The focus of this thesis is on lung cancer, so I chose to use in-house datasets collected from the Oxford University Hospitals (OUH) and hospitals participating in the DART lung health programme [81]. Both **OUH** lung and **DART** lung datasets are described in detail in Section 3.2.3. OUH dataset came in 3 batches with the first batch scanned at 20x and at 40x on different scanners. The DART dataset came from 4 different DART sites, which we denote DART_001 - DART_004.

I also utilised the publicly available lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC or LSCC) collections from The Cancer Genome Atlas (**TCGA**) [4] and The Cancer Imaging Archive - Clinical Proteomic Tumor Analysis Consortium (**TCIA-CPTAC**) [8, 9]. Both have been extensively used to train and evaluate novel computational pathology algorithms.

TCGA datasets [4] contain imaging and genomics data from different types of cancer. The lung portion contains 541 digitised slides from 478 Lung Adenocarcinoma (LUAD) cases and 512 digitised slides from 478 Lung Squamous Cell Carcinoma (LUSC) cases.

TCIA-CPTAC datasets [8, 9] contain whole slide images, radiology, and proteomics data from different human cancers. The lung portion contains slides with lung adenocarcinoma (LUAD - 591 slides), lung squamous cell carcinoma (LUSC or LSCC - 519 slides), and non-cancerous tissues (706 slides). All of the non-cancerous slides come from patients who have one of the two cancers (LUAD or LUSC). There are no patients who did not have lung cancer in the lung portion of the TCIA-CPTAC dataset. Some of the LUAD slides have labels for the presence of the adenocarcinoma patterns.

To further validate the hypothesis mentioned above, I chose the **CAMELYON16** [122] dataset of breast cancer tissue commonly used for the evaluation of new computational pathology algorithms. This dataset has been released as part of the 2016 ISBI challenge on cancer metastasis detection in sentinel lymph nodes and includes 399 slides from Radboud University Medical centre and University Medical centre Utrecht with 269 images

(159 normal, 110 metastasis) in the official training set and 130 images (80 normal, 50 metastasis) in the test set that has been subsequently released.

Slide Format. I now need to emphasise the fact that the pathology slides were being scanned using different scanners at different resolutions (measured in microns per pixel, which we will denote μpp) and in different slide formats. OUH and DART slides have been scanned using Hamamatsu (0.45 microns per pixel, ".ndpi" format) and Ventana DP200 (20x: 0.5 microns per pixel, 40x: 0.25 microns per pixel) scanners. TCGA-lung and TCIA-CPTAC lung slides have been scanned in various resolution/power configurations (20x: 0.25 μpp, 20x: 0.5 μpp, 40x: 0.25 μpp, and 40x: 0.5 μpp,) and saved in ".svs" format. CAMELYON16 slides have been scanned at 40x magnification (0.23 microns per pixel) and saved in ".tif" format.

Standard Model Input. Many of the released pathology foundation models have been trained on 224x224 pixels patches extracted from H&E slides at 0.5 microns per pixel, which is agreed upon, but not enforced, standard resolution for a 20x magnification setting. Hence, these models also expect to receive patches at the same resolution. To deal with multiple slide formats, pyramidal levels, and scanning resolutions, I utilised one of the more developed digital pathology libraries called TIAToolbox [145]. To use it on the non-tiled ".tif" images produced by the Ventana DP200 scanner used for the DART project, I contributed code to the TIAToolbox library [3].

4.4.2 WSI Datasets: Evaluation of Augmented Patch Similarities

To efficiently select a promising foundation model for a new dataset, I created a pretext task that follows the evaluation procedure I proposed for foundation models on the LC25000-clean dataset (Section 4.3.1). I (1) selected a small number of patches from each of the slides of the dataset, (2) produced augmented versions of these patches, (3) computed their embeddings using the candidate foundation models, and (4) evaluated

[3] https://github.com/TissueImageAnalytics/tiatoolbox/pull/807

how well a K-Means algorithm with a cosine similarity metric can separate the clusters of augmented patches from different origins. I followed the standard approach of using raw features produced by the models since any form of post-processing might favour one of the models and skew the results.

Patch Sampling and Augmentation. The sampling and augmentation process was done as follows. First, I produced a tissue mask using Otsu thresholding using the TIAToolbox [145] library. Then, I sampled 50 patches of 448x448 pixels at 0.5 microns per pixel resolution (standard scanning protocol for many types of cancers and, hence, the standard for many foundation models). I only considered patches with at least 50% of tissue since the goal was to evaluate how well the models can extract features from tissue and not from the background. After that, I made 10 rotational augmentations spread uniformly (at 0, 36, 72, ..., 324) degrees and centre-cropped the central 224x224 region, the standard size for model input. Sampling regions larger than the standard model input allows to avoid the distortions that come from zooming into the central part of the image. Figure 4.1 shows these distortions: the right-most image is zoomed out compared to the other four images. This resulted in 500 patches per slide. This gave a 10x speed-up in feature computation compared to a 10000-patch slide, in which half of the patches (5000) have tissue. It is known that these 500 patches came from 50 clusters of augmented images, which means that out of $500 \times 499/2 = 124750$ possible connections, $50 \times (10 \times 9)/2 = 2250$ are genuine connections the clustering algorithm should find. Other augmentations could also be considered—most notably, stain augmentation, which involves adjusting the colour distribution of one image to match that of another, simulating variability in staining protocols across different hospitals. Since stain augmentation has been shown to be one of the most effective ways to improve training and help models generalise to unseen datasets [16, 152, 172, 201], I believe it could also be used to assess whether the features extracted by a foundation model will generalise well across the entire dataset of interest.

Feature Extraction. I chose 6 feature extractors: 3 trained on public and 3 on proprietary datasets. To compute the features on the augmented patches, I used the normalisation procedure advised by the model authors of each feature extraction model listed below as "Model name. Training data. Input normalisation. Output dimension.".

1. ResNet18-lung [117]. TCGA-LUAD, -LUSC (1040 slides). No norm. 512 dim.
2. ResNet18-CAMELYON16 [117]. CAMELYON16 (399 slides). No norm. 512 dim.
3. Phikon-v2 [74]. Public datasets. (58,400 slides); ImageNet norm. 1024 dim.
4. UNI [46]. Closed dataset of 100,000 slides. ImageNet norm. 1024 dim.
5. Prov-GigaPath [195]. Closed dataset of 170,000 slides. ImageNet norm. 1536 dim.
6. Virchow-v1 [185]. Closed dataset of 1,500,000 slides. ImageNet norm. 2560 dim.

The ResNet18 feature extractor models were trained on specific cancer types by Li et al. [117], so I used the TCGA-lung model for all lung datasets (OUH lung, DART lung, TCGA lung, TCIA-CPTAC lung) and the CAMELYON16-trained model for the CAMELYON16 dataset of breast tissue. Phikon-v2 [74] was trained on public data, including TCGA and TCIA-CPTAC, but not on the CAMELYON16 dataset. Due to this data leakage, we can expect particularly good performance of ResNet18-lung on TCGA-lung, ResNet18-CAMELYON16 on CAMELYON16, and Phikon-v2 on TCGA-lung and TCIA-CPTAC lung datasets.

Clustering Performance Evaluation. I computed the same metrics as for the LC25000-clean benchmark described in more detail in Section 4.3.1: binary connectivity metrics (Accuracy, Precision, Recall, F1-score, Specificity, Balanced Accuracy) and clustering metrics (Fowlkes–Mallows index [75], Adjusted Rand Index [98, 150], Normalized Mutual Information score, Homogeneity, Completeness, V-Measure [154]), clustering-agnostic metrics (precision@1 and precision@5).

Figures 4.6 and 4.7 show the performance of the models on all our datasets. The general trend shows that the Prov-GigaPath [195] model performed best, followed by UNI [46], then the ResNet18 models trained on TCGA-lung [117] (evaluated on all lung datasets) and CAMELYON16 (evaluated only on CAMELYON16) [117]. The worst results are shown by Phikon-v2 [74] and Virchow-v1 [185] feature extractors.

Although CAMELYON16-trained ResNet18 outperformed UNI on the CAMELYON16 dataset (possibly due to data leakage during pre-training), TCGA-lung-trained ResNet18 does not outperform UNI. This could be attributed to a much more homogenous nature of the CAMELYON16 (scanning resolution) when compared to TCGA-lung that could have reduced the quality of pre-training by Li et al. [117].

On the radar plot (Figure 4.6, top), I split the OUH and DART datasets into their respective parts when I report the Fowlkes-Mallows Index. A drop can be seen in Virchow-v1 performance on OUH batch 1 images scanned with the Hamamatsu scanner at 20x compared to the OUH batch 1 images scanned with the Roche Ventana DP200 scanner at 40x. Because all other models exhibited relatively stable performance, this can be attributed to the smaller prevalence of the specific features of the Hamamatsu scanner in the Virchow-v2 training dataset.

The parallel coordinates plots (Figures 4.6 and 4.7) show that (1) Accuracy, Specificity, and Precision@1 will not be useful to rank the models as candidates for downstream tasks since they collapse near 1 for all models and datasets (2) there is no difference in the ranking produced by other metrics, and (3) that Precision, F1 Score, Fowlkes-Mallows Index, and Adjusted Rand Index show the most separation between the lines and hence are most suited to be used when measuring the Pearson correlation coefficient between the clustering and downstream classification performance.

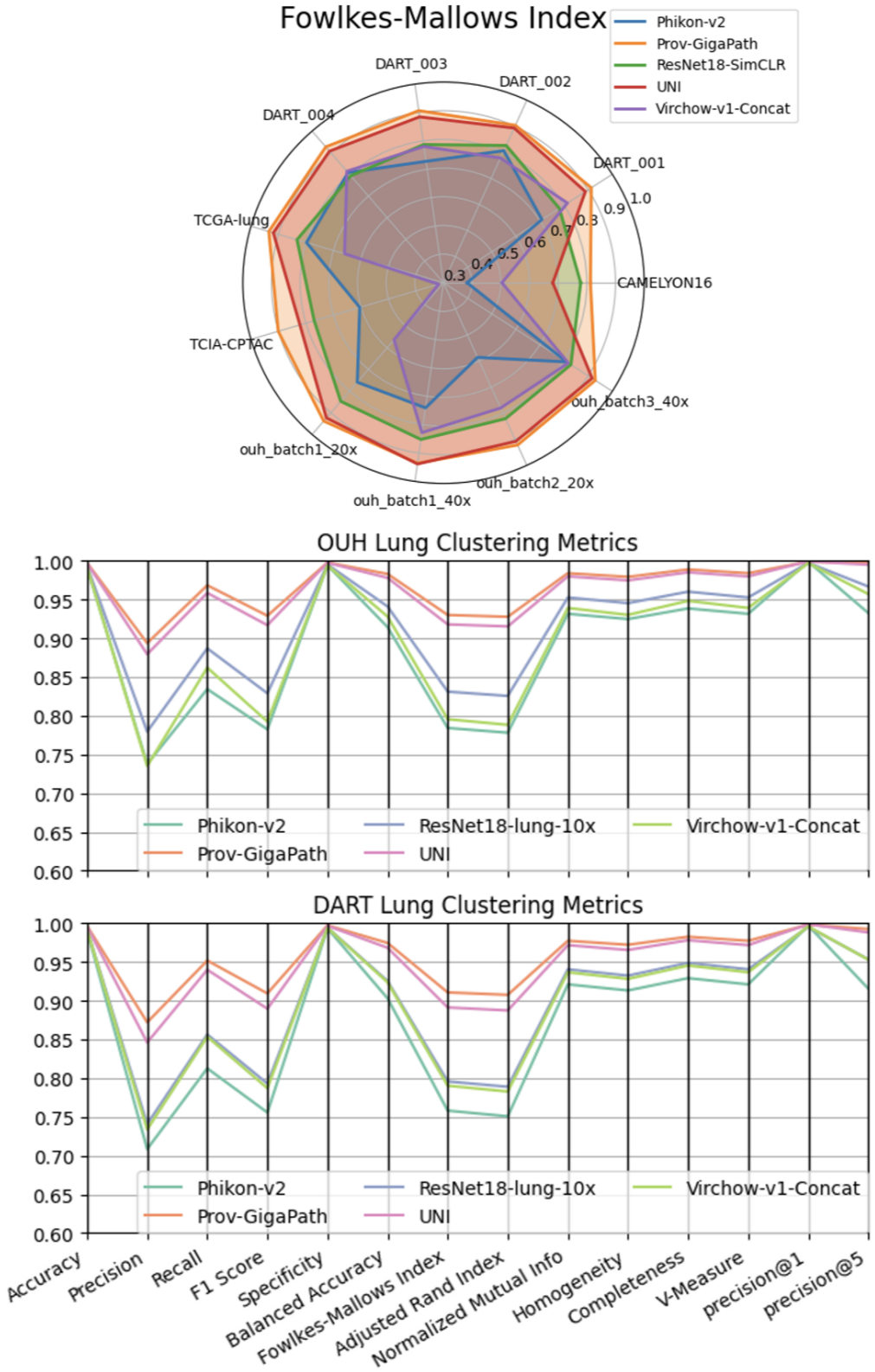

Figure 4.6: Clustering performance. Top: FMI for all datasets. Performance was reported separately for different parts of the OUH and DART datasets. Performances of ResNet18 trained on TCGA-lung and CAMELYON16 [117] are merged into one line (ResNet18-SimCLR) with the CAMELYON16 model only used on the CAMELYON16 dataset. Middle and bottom: all recorded slide metrics aggregated for OUH and DART lung datasets.

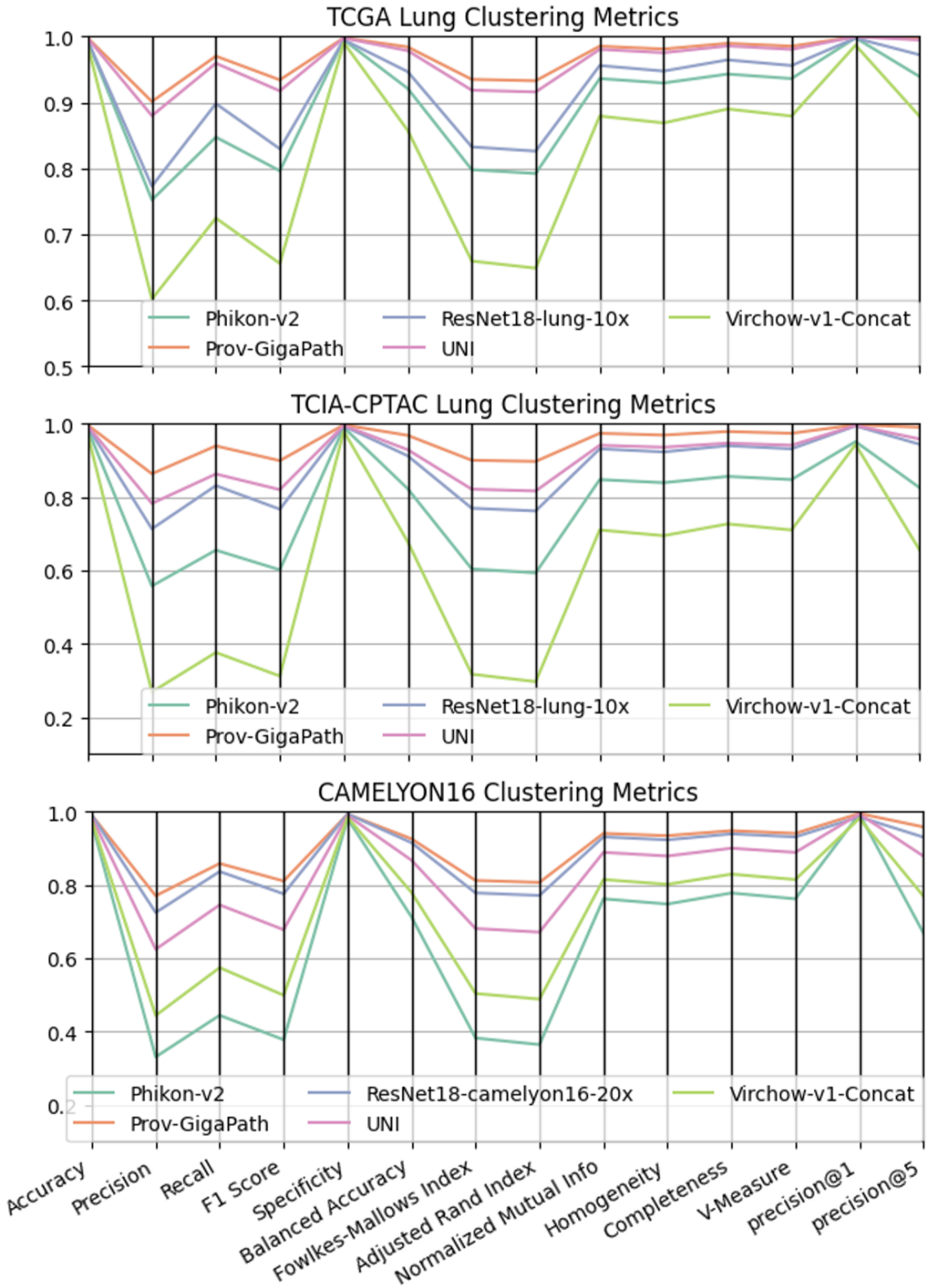

Figure 4.7: Clustering performance: all recorded slide metrics aggregated for TCGA lung (top), TCIA-CPTAC lung (middle), and CAMELYON16 breast (bottom) cancer datasets.

4.4.3 WSI Datasets: Classification with AB-MIL

To check if the pretext task of clustering augmented patches can approximate the performance ranks of foundation models when used for downstream tasks, I gathered the results obtained in published studies (on TCIA CPTAC lung and CAMELYON16 breast) and performed the downstream tasks myself on TCGA lung (LUAD vs LUSC binary classification) and combined OUH+DART lung (LUAD vs Other binary classification) datasets. All public datasets have been described in detail in Section 4.4.1, while OUH and DART lung datasets are described in Section 3.2.3. Although this work focuses on choosing and evaluating foundation models in the context of lung cancer, I included the CAMELYON16 dataset to show that this clustering evaluation method can be used to choose models outside the lung cancer domain.

To aggregate the patch features produced by the foundation models into classification predictions during downstream classification, I report the results for AB-MIL [102] model both where I extract results from papers and where I trained downstream classifiers myself. I chose the AB-MIL model out of all aggregators because it has passed the test of time since its release in 2018 by showing competitive results when used on top of a strong feature extractor [36, 48, 137, 191].

I benchmarked ResNet18-lung [117], Phikon-v2 [74], UNI [46], Prov-GigaPath [195], and Virchow-v1 [185]. These models' training data and normalisation details are described in Section 4.4.2. Li et al. [117] pre-trained two versions of ResNet18, one on TCGA lung data and the other on CAMELYON16 data. So, I used the ResNet18 TCGA-lung model for all lung datasets and reported results for ResNet18 CAMELYON16 model on CAMELYON16 data. To train Phikon-v2, Filiot et al. [74] used multi-organ cohorts from many public datasets, including TCGA and TCIA-CPTAC. As mentioned before, pre-training and evaluating a model on the same dataset can result in inflated performance due to data leakage. Nevertheless, I included ResNet18 TCGA-lung and CAMELYON16 models [117] and Phikon-v2 [74] to compare the effect of leaking the data against using

larger proprietary datasets for pre-training UNI [46], Prov-GigaPath [195], and Virchow-v1 [185].

Table 4.3 shows downstream classification performance reported for each extractor model in papers. I could not find the benchmarking results for TCGA lung for most models, so I evaluated the model performance on TCGA lung myself (Table 4.4). Observations on the classification performance and the agreement with clustering results are detailed in Section 4.4.4 and summarised in Table 4.2, respectively.

Training Data Split

For TCGA lung, I used the split into train+validation (831 slides) and test (215 slides) sets provided by Li et al. [117] to contain the data leakage of ResNet18 [117] to the training set. I uses a stratified group split for the combined OUH and DART lung datasets, ensuring all slides from a single patient end up together in training, validation, or test sets. I also stratified based on the LUAD label. For both datasets, the data was split in an 80/20 ratio into train+validation / test split. The validation set was then obtained by taking 20% from the train+validation data, resulting in 64/16/20 proportions.

Training Details

In my own experiments, I used the open-source TIAToolbox library [145] to produce tissue masks using Otsu thresholding and extract features at 0.5 microns per pixel (standard for most foundation models) from all patches that had at least 10% tissue detected (pre-processing pipeline from Prov-GigaPath [195]). Figure 4.8 shows the distribution of patches per slide in TCIA-CPTAC lung (top-left), CAMELYON16 (top-right), TCGA lung (bottom-left), OUH+DART (bottom-right) datasets. TCGA has 21% of slides with fewer than 5000 patches, while this proportion is between 41% and 46% for TCIA-CPTAC lung, CAMELYON16, and OUH+DART datasets.

Inspired by the strong benchmarking results of Filiot et al. [74], I followed a similar train-

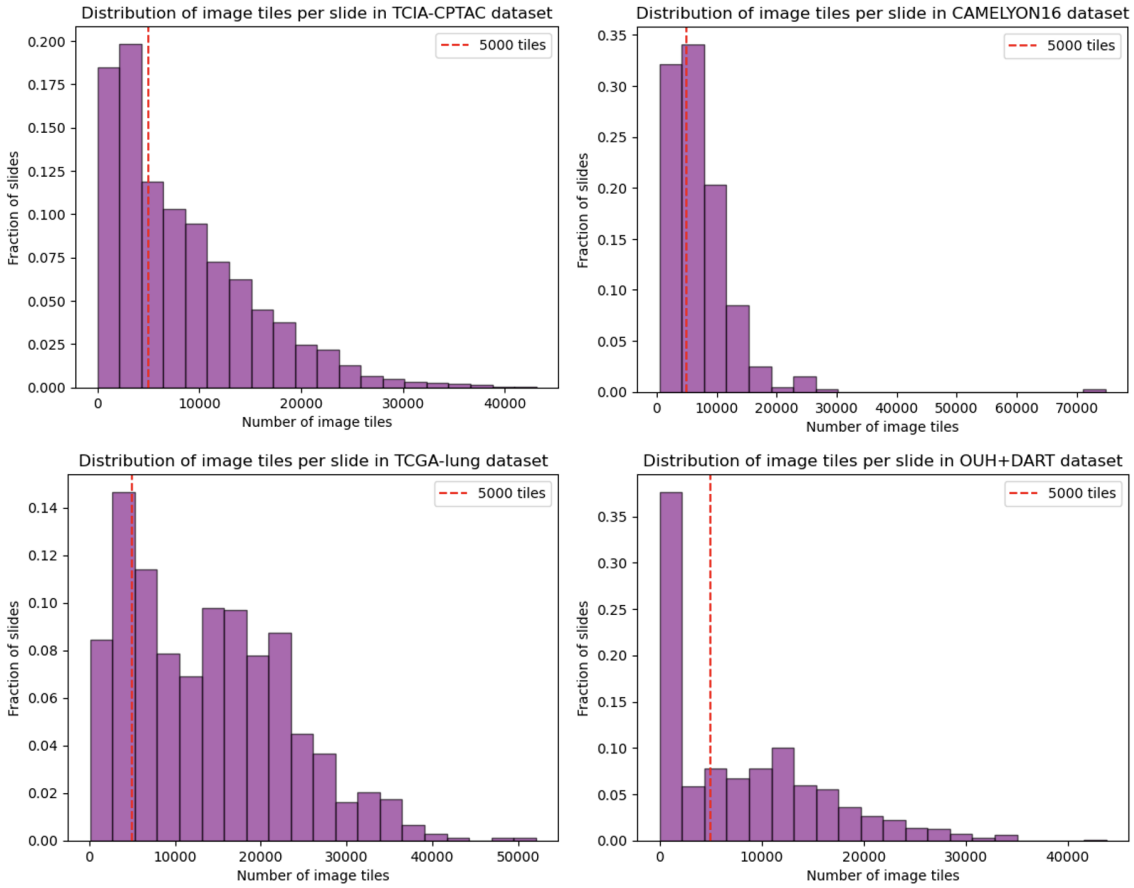

Figure 4.8: Distribution of tiles per slide in TCIA-CPTAC lung (top-left), CAMELYON16 (top-right), TCGA lung (bottom-left), OUH+DART (bottom-right) datasets. The red vertical dashed line indicates 5000 patches sampled from slides with more than 5000 patches at every training iteration.

ing pipeline. At each epoch, I randomly sampled 5000 patches per slide (for slides containing more than 5000 patches - Figure 4.8) and constructed batches of size 16, padding as necessary to obtain input tensors of shape $(16, 5000, \text{embedding size})$. In addition to improving training efficiency by ensuring uniform batch size, this random sampling acted as a form of regularisation. Similar in spirit to Dropout [164] or random cutout augmentation [66] used for natural image models, subsampling encouraged the MIL aggregator to avoid over-relying on a small set of highly informative patches. Instead, the model was forced to identify useful patterns across varying subsets of patches (different every epoch), helping it to learn less obvious dependencies and escape local minima.

I used a standard single-branch AB-MIL architecture with gated attention and a projection

size of 128. Before passing the patch-level embeddings to the AB-MIL aggregator, I reduced the feature dimensionality to 128 using a fully connected layer. The resulting bag embedding was fed into a fully connected binary classification layer. Training was performed for 100 epochs using the Adam optimiser [110] with a constant learning rate of $1e{-}3$, β values of (0.9, 0.999), and no weight decay.

During evaluation, I did not apply patch subsampling, but instead used all extracted patches to present the full slide content to the model. This approach ensures that no diagnostically relevant tissue is omitted at test time, allowing the MIL aggregator to operate on the complete information available and produce the most comprehensive and reliable prediction for each slide. Since training involves random subsampling, using all patches at evaluation time also reduces prediction variance and provides a more stable estimate of model performance. However, this approach does not guarantee improved accuracy, as it introduces a distribution shift between training and evaluation: the model is exposed to a different patch distribution at test time than during training. Nonetheless, a full-patch evaluation remains a common strategy for obtaining stable and reproducible results when running inference with MIL models.

Alternatively, if inference time and computational cost are not a concern, one could perform multiple forward passes with different random patch subsamples at evaluation time, analogous to MC-Dropout [76], and average the resulting slide-level predictions. This would treat patch sampling as a source of uncertainty and yield predictions with a measure of uncertainty, but at a significantly higher inference cost.

I monitored the AUC on the validation set to select the best checkpoint, which was then used to obtain the final test results reported in Table 4.4. I experimented with the training set-up in the following ways: did not subsample patches (used a smaller batch size to fit into memory), set different learning rates (0.001, 0.0005, 0.0002, 0.0001), used learning rate schedulers, varied the downsampling factor from the extractor embedding size to AB-MIL's input (constant: 128, 256, 512, and dependent on input embedding size), and

used an AB-MIL version with non-gated attention. However, the training pipeline used by Filiot et al. [74] resulted in the best performance for each of the models presented in Table 4.4.

4.4.4 WSI Datasets: Agreement between Clustering & Classification

For each dataset and foundation model combination, I recorded the clustering (Figures 4.6, 4.7) and downstream classification performance (Tables 4.3, 4.4). The combined comparison of clustering and classification performance ranks is shown in Table 4.2.

Model	**CAM16**		**TCIA-CPTAC**		**TCGA-lung**		**OUH+DART**	
	Cluster	Class	Cluster	Class	Cluster	Class	Cluster	Class
ResNet18 [117]	3	5	3	–	3	1	3	1
Virchow [185]	4	1	5	1	5	4	4	4
Phikon-v2 [74]	5	1	4	–	4	2	5	2
UNI [46]	2	1	2	1	2	3	2	3
GigaPath [195]	1	1	1	1	1	5	1	5

Table 4.2: Clustering and classification rankings (1 = best, 5 = worst) of feature extractors across four datasets. "CAM16" refers to the CAMELYON16 dataset. "Virchow" and "GigaPath" are abbreviations for Virchow-v1-Concat and Prov-GigaPath, respectively. Classification rankings for CAM16 and TCIA-CPTAC are from published papers, while TCGA-lung and OUH+DART results are from my experiments. Dashes (–) indicate classification results missing in published literature. Rankings are based on Fowlkes–Mallows index [75] for clustering and on the AUC scores of AB-MIL classifiers. Identical ranks indicate ties.

Results from Published Works

I did not run classification experiments on the CAMELYON16 and TCIA-CPTAC lung datasets. Instead, I report the results presented in [74, 117, 137] in Table 4.3.

On CAMELYON16, the clustering performance ranked as follows (best to worst): Prov-GigaPath, UNI, ResNet18 (trained on a subset of CAMELYON16 by Li et al. [117]), Virchow-v1-Concat, and Phikon-v2. For the Metastasis vs. Normal classification task, Filiot et al. [74] reported AUC scores around 0.99 for UNI, Phikon-v2, Prov-GigaPath, and Virchow-v1-Concat, while Li et al. [117] reported an AUC of 0.9 for ResNet18, pre-trained on CAMELYON16 using SimCLR [49].

	TCGA	TCIA-CPTAC	CAM16
ResNet18 [117]	0.9488[117]	-	0.8653[117]
Virchow-v1 [185]	train for [137] ->	0.98[137]	0.989[74]
‡**Phikon v1** [73]	train for [73], [137] ->	0.95[137]	1.000[74]
Phikon-v2 [74]	train for [74]	train for [74]	0.997[74]
UNI [46]	train for [137] ->	0.99[137]	0.998[74]
Prov-GigaPath [195]	0.756[195]; train for [137] ->	0.98[137]	0.995[74]

Table 4.3: Reported Classification performance (AUC) of AB-MIL classifiers for TCGA lung (LUAD vs LUSC subtyping), TCIA-CPTAC lung (LUAD vs LUSC subtyping), and CAMELYON16 (normal vs metastasis subtyping) datasets in papers. The subscript to the right of the AUC indicates the source paper where the results come from. Filiot et al. [74] used both TCGA and TCIA-CPTAC datasets for training Phikon v2, so neither Filiot et al. [74], nor Neidlinger et al. [137] report Phikon-v2 results on these datasets. ‡ Phikon v1 (trained on TCGA data) were added to the table as a baseline in the absence of a score of Phikon-v2 on TCIA-CPTAC; they have not been evaluated in a clustering benchmark in Section 4.4.2. Evaluation by Filiot et al. [74] included ensembling of 5 best AB-MIL models trained with different initialisations, so the results for the CAMELYON16 dataset are inflated because of that. Neidlinger et al. [137] trained the aggregator models on the TCGA lung cohort ("train for [137] ->") and evaluated them on the TCIA-CPTAC lung cohort hence they only reported performance on TCIA-CPTAC data.

Two observations follow from these results: (1) while Virchow-v1-Concat and Phikon-v2 features were insufficient for clustering, they achieved near-perfect AUC in classification; (2) ResNet18 performed better than Virchow-v1-Concat and Phikon-v2 in clustering, but worse in classification. One possible explanation is that Filiot et al. [74] recorded multiple checkpoints and ensembled models with the best validation AUC, while Li et al. [117] relied on a single model and may have under-optimised after achieving a state-of-the-art result. Although Filiot et al. [74] did not mention stain normalization, the absence of publicly available code means that a well-chosen training-time stain augmentation and/or test-time normalization could have contributed to their high performance on CAMELYON16.

On the TCIA-CPTAC lung dataset, the clustering ranking was: Prov-GigaPath, UNI, ResNet18 (trained on a subset of TCGA-lung by Li et al. [117]), Phikon-v2, and Virchow-v1-Concat. For the LUAD vs. LUSC classification task, Neidlinger et al. [137] reported AUC scores of 0.98–0.99 for UNI, Prov-GigaPath, and Virchow-v1-Concat. Li et al. [117] did not evaluate on TCIA-CPTAC, and, to the best of my knowledge, no published

results exist for Phikon-v2 on this dataset, as it was used for training the Phikon model. However, the 0.95 AUC reported for Phikon-v1 by Neidlinger et al. [137] suggests that Phikon pre-training enabled extraction of useful features for lung subtype classification in TCIA-CPTAC.

Similar to CAMELYON16, I observed that Virchow-v1-Concat did not perform well in clustering but supported training of a highly accurate AB-MIL classifier on TCIA-CPTAC. Neidlinger et al. [137] did not report the image normalization strategy, but the original STAMP repository supports Macenko normalization[4], which may have contributed to their strong classification results and was not used in my clustering setup.

To conclude, comparisons of clustering and classification performance reported in the literature on CAMELYON16 and TCIA-CPTAC lung datasets do not show a consistent relationship between clustering quality and the ability to train an accurate downstream classifier.

Results from My Experiments

With only two reported performances on the TCGA-lung dataset (ResNet18: 0.9488 [117], Prov-GigaPath: 0.756 [195]) and no published results for the in-house OUH+DART dataset, I performed the experiments myself reporting the results in Table 4.4.

TCGA-lung. The clustering rankings on TCGA-lung were (best to worst): Prov-GigaPath, UNI, ResNet18 (trained on part of TCGA-lung by Li et al. [117]), Phikon-v2, and Virchow-v1-Concat.

For classification, I was able to match the literature-reported AUC of >0.9 with ResNet18-TCGA-lung [117]. Phikon-v2, UNI, and Virchow-v1-Concat achieved AUCs between 0.75 and 0.78. In contrast, I was unable to train an accurate classifier using Prov-GigaPath embeddings (AUC: 0.558).

ResNet18's strong performance is unsurprising—it was trained specifically on TCGA-

[4] https://github.com/KatherLab/STAMP/issues/42

	TCGA lung	OUH+DART lung
†**ResNet18** [117]	**0.927**	**0.984**
Virchow-v1 [185]	0.754	0.714
†**Phikon-v2** [74]	0.778	0.824
UNI [46]	0.775	0.820
Prov-GigaPath [195]	0.558	0.649

Table 4.4: Best classification performance (AUC) of AB-MIL classifiers for TCGA lung (LUAD vs LUSC subtyping) and OUH+DART (LUAD vs rest subtyping) datasets. TCGA-lung was used by Li et al. [117] to pre-train †ResNet18 and by Filiot et al. [74] as one of the public training datasets for †Phikon v2.

lung data and reached 0.9 AUC within just 10 epochs. Phikon-v2 and UNI (embedding size 1024) required around 30 epochs, while Virchow-v1-Concat (embedding size 2560) converged at 55 epochs with 0.75 AUC. This suggests that larger embeddings can slow convergence and that Phikon's exposure to TCGA-lung during training did not lead to better downstream performance than UNI or Virchow-v1-Concat.

Although Prov-GigaPath was trained with 45% of its data from lung cancer slides [195], its poor classification performance might also be attributed to domain shift from staining or scanning differences between the training data and TCGA. Its embedding size of 1536 alone does not explain the poor results—Virchow-v1-Concat's larger embedding size (2560) still yielded better results. I also considered whether subsampling 5000 patches during training caused underfitting (due to changing subsets per epoch), but removing the subsampling did not improve performance. Furthermore, TCGA-lung slides are known to contain on average 80% tumour tissue, and training a patch-level classifier assuming all patches share the slide label can already achieve 0.9 accuracy [117]. Another possible reason for the poor performance of Prov-GigaPath is underfitting caused by poorly chosen hyperparameters, which could be resolved through a more extensive hyperparameter search.

OUH+DART. Clustering rankings were (best to worst): Prov-GigaPath, UNI, ResNet18 (trained on a part of TCGA-lung by Li et al. [117]), Virchow-v1-Concat, and Phikon-v2. Clustering performance on the OUH+DART dataset was higher than on TCGA-lung,

TCIA-CPTAC lung, or CAMELYON16.

The best classification performance was again achieved by ResNet18-TCGA-lung (AUC: 0.984), followed by UNI and Phikon-v2 (AUCs $\sim$0.82), then Virchow-v1-Concat (0.714), and finally Prov-GigaPath (0.649). As in TCGA-lung, ResNet18 remained the best performer, with UNI and Phikon-v2 again showing solid performance. Unlike in TCGA-lung, however, the OUH+DART data had not been seen by either ResNet18 or Phikon-v2. Their strong results suggest a similarity in distribution between OUH+DART and the TCGA-lung dataset, rather than data leakage. Conversely, Prov-GigaPath's poor classification performance could be explained by a stark distribution shift between it's training data and OUH+DART or underfitting as has already mentioned when describing the results for TCGA-lung classification.

All together, these experiments show that the best features for capturing rotational invariance, and consequently clustering augmented patches (Prov-GigaPath), can yield the worst classification performance when used with a simple AB-MIL aggregator. In contrast, features from a smaller model like ResNet18, pre-trained on a dataset similar in distribution to the target dataset, can be highly effective for downstream classification.

4.5 Conclusions

Data leakage threatens the utility of many valuable histopathology datasets, leading to models that overfit and exhibit artificially high performance that does not reflect their true capabilities. My work addresses this issue with the LC25000 dataset of histology tiles. By developing a semi-automatic pipeline, I effectively curated the LC25000 dataset, ensuring it is free from data leakage and can be utilised to profile machine learning models more accurately. The cleaned dataset and the associated pipeline are released to the research community.

Furthermore, this study shows that grouping augmented images from the same origin can

serve as a minimal setup benchmark for evaluating the quality of tile-level features learned by different vision models when used for a subsequent **tile classification** task. Pathology-specific vision models outperformed general natural image models. This superiority is evident even in baseline clustering tasks, where ImageNet pre-trained models exhibit a considerable error rate. This highlights the need for foundation models trained specifically to learn histopathology features.

Finally, I observed that the ability of pathology-trained models to cluster patches extracted from a new dataset and augmented using rotations, flips, and crops **can not be used** to reliably choose the best model for downstream **slide classification** and further research in this direction is required. Adding stain augmentation can potentially lead to more reliable results by requiring the models to exhibit stain invariance in addition to the rotational invariance.

Concurrent work "Are the Latent Representations of Foundation Models for Pathology Invariant to Rotation?" [69] was published in December 2024. This work focuses on similar questions to Section 4.4.2, but does not intersect with other sections of this chapter.

Chapter 5

Subtyping Lung Cancers

Contents

Contributions

Dependency-MIL [33]

- Constructed a new dataset with lung cancer-associated subtype and pattern labels combined with datasets from publicly available repositories: TCGA, TCIA-CPTAC, and DHMC.
- Proposed an approach to model class dependencies between cancer subtypes and histological patterns in a weakly-supervised multi-label setting in the presence of partial labels.
- Achieved subset accuracy of 84% when differentiating lung cancer subtypes and cancer-associated histological patterns.
- Released compiled labels for the public part of the dataset and full code for tissue segmentation, feature extraction, training, and evaluation: `https://github.com/GeorgeBatch/dependency-mil`

DART Data, Foundation Models, Mixed Supervision

- Evaluated the best Dependency-MIL model on newly collected DART data.
- Benchmarked pathology foundation models on the proposed multi-label lung cancer classification task.
- Showed that adding minimal region annotation for mixed supervision improves classification performance.

This chapter includes my first author papers "Accurate subtyping of lung cancers by modelling class dependencies" [33] presented at the International Symposium on Biomedical Imaging (ISBI) 2024 conference and "38 Modelling Class Dependencies for Lung Cancer Subtyping from Digitised Pathology Images" [32] presented at the British Thoracic Oncology Group (BTOG) 2024 conference. Together with a more extensive evaluation of the DART data and work on focusing on diagnostic regions, they form Chapter 5.

5.1 Introduction

Lung cancer constitutes the primary cause of oncological mortality worldwide [177] with three main subtypes: Non-small Cell Lung Carcinoma (NSCLC), Small Cell Carcinoma, and Carcinoid Tumour. Non-Small Cell Carcinomas, represented by Squamous Cell Carcinomas (LUSC) and Adenocarcinomas (LUAD) account for more than 80% of all lung cancer cases [60, 95]. Adenocarcinomas, representing around 50% of all cases [133], are categorised into sub-classes by the presence of adenocarcinoma-specific morphological patterns: acinar, lepidic, micropapillary, papillary, solid. The cancer subtype and associated oncological histomorphology are crucial indicators for cancer prognosis and treatment [140]. Hence it is vital to identify both the broader classes and the pattern-determined sub-classes within them.

Histopathological subclassification of lung adenocarcinoma patterns is particularly challenging since these patterns are often found in combination within the same tumour. Existing publicly available datasets [190] classify LUAD according to the pattern most represented in the cross-sectional area of the histological section (so-called predominant pattern). However, several studies have revealed that secondary histological patterns can be useful in the management of multiple LUAD, as their morphological features can be applied to support the diagnosis of multiple synchronous or asynchronous cancers [112]. Therefore, in addition to identifying the subtypes of cancer, this work aims to determine all the histological patterns associated with LUAD.

5.2 Literature

Progress has been made towards computationally subtyping lung cancers from H&E whole slide images (WSIs). Due to the difficulty of obtaining clinical samples with annotations and fine-grained pattern labels, many works focused on classifying weakly-labelled (slide-level labelled) slides from public databases such as TCGA and TCIA-CPTAC [10] lung cancer cohorts into LUAD, LUSC, and benign tissue [55, 117, 124]. Other groups, based on extensive region-based annotations, trained models to predict the predominant morphological pattern of LUAD from patches [190, 20]. Alsubaie et al. [20] validated their approach on patches from LUAD slides, while Wei et al. [190] used a sliding-window-based method and heuristically aggregated patch-level predictions into slide-level predictions on a test dataset from Dartmouth Hitchcock Medical centre (DHMC). Qiao et al. [147] were the first to consider both lung cancer subtyping and LUAD pattern prediction on the TCGA and TCIA-CPTAC [10] patches. First, a LUAD/LUSC classifier was trained on TCGA patches. Then, separate classifiers were fine-tuned on patches obtained from extensive region-based annotations akin to [190, 20] from in-house slides and TCIA-CPTAC [10] to identify tumour-associated patterns on LUAD patches (separate classifier for each pattern). These patch-level pattern classifiers were combined using bootstrap aggregation (bagging) strategy [37] into a pattern prediction classifier.

Popular Multiple-Instance Learning (MIL)-based methods for learning from weak labels (AB-MIL [102], DSMIL [117], CLAM [124]) showed the effectiveness of using separate attention branches for different classes but allow little to no communication between class-specific embeddings before the classification layer. This prevents the models from modelling dependencies between the classes, which might not pose problems in binary or multi-class classification tasks but can hinder performance when class dependencies are present in a multi-label scenario.

Some researchers used a combination of weak slide-level labels and tile-level annotations to improve classification performance. Tourniaire et al. [178] added mixed supervision to

AB-MIL [102] by fitting an instance classifier on top of the additional instance embedder layer to predict one of the two label classes (normal vs metastasis) on the CAMELYON16 [122] dataset. The exhaustive annotations of tumour regions available for CAMELYON16 were used as patch labels. In a follow-up work Tourniaire et al. [179] extended the CLAM framework [124] by adding explicit supervision of the attention weights again using the CAMELYON16 [122] slides and annotations. In both works, all slides were classified as either normal or metastasis, and all patches had one of these two labels.

5.3 Contributions

Unlike earlier patch-level studies [190, 20, 147], I proposed to learn the subtyping and presence of LUAD patterns in a weakly supervised manner without relying solely on patch-level labels, instead using only slide-level labels that indicated cancer subtypes or the presence of adenocarcinoma patterns.

In contrast to prior slide-level studies [102, 117, 124], I facilitated the learning of class interactions by incorporating a class-communicator module within the network. Furthermore, I framed the task as a multi-label problem (Figure 5.1), where the sample labels comprised a series of binary indicators denoting the presence of cancer subtypes and/or adenocarcinoma patterns. This approach was motivated by the understanding that (1) pathologists diagnosed LUAD based on the occurrence of adenocarcinoma patterns, (2) benign and LUSC regions lacked any adenocarcinoma patterns, (3) multiple tumour subtypes and/or adenocarcinoma patterns could coexist within the same slide, and (4) some indicators in the labels might be absent due to partial observation during pathological examination.

I extended this method to utilise non-exhaustive selections of diagnostic and benign regions chosen by pathologists while subtyping in-house OUH and DART slides according to the DART Annotation Protocol (Section 3.2.2). Region selection did not increase the subtyping time, as pathologists were asked to mark only the diagnostic regions they

used for subtyping, as well as a few benign tissue regions if they could identify them quickly. Unlike the CAMELYON16 annotation used by [178] for mixed supervision, our region selection process was non-exhaustive, meaning I performed binary classification (diagnostic vs benign) with partial labels for slides containing tumours and full labels for benign slides, since, by definition, none of the patches on benign slides exhibited tumour features. Furthermore, the region labels did not directly correspond to the multi-label scenario. Although diagnostic regions could be used directly for LUAD and LUSC labels, they could not serve as explicit supervision for LUAD patterns, which could be present on some diagnostic regions of LUAD slides and absent from others.

The work presented in this chapter was conducted over 1.5 years, during which I gained access to more data. Moreover, several computational pathology extractors were released in the final year of my doctoral project. My contributions can be categorised into two groups. The first group, published as the ISBI 2024 conference paper [33], centred on developing a new MIL aggregation method within a weakly supervised partial-label setting focused on a multi-label classification scenario. The second group provided a more comprehensive evaluation of various feature extractors (including foundation models) on a larger dataset alongside a mixed-supervision method aimed at concentrating on diagnostically relevant patches in the presence of partial labels.

5.3.1 Contributions: Dependency-MIL with Weak Supervision

First, I constructed a multi-label dataset combining TCGA, TCIA-CPTAC, and DHMC [190] datasets with only slide-level labels. The LUAD pattern labels were either parsed from the TCIA-CPTAC cohort-information document [10] or available alongside the DHMC dataset. I incorporated samples from the Oxford University Hospitals (OUH) with slide-level labels specifying the presence of cancer subtypes and LUAD patterns. Second, I proposed **Dependency-MIL** - a method for modelling class-dependencies that allowed the learning of robust bag representations suitable for multi-label problems under weakly-supervised conditions with partial labels (see Figure 5.1). This addressed the gap left by

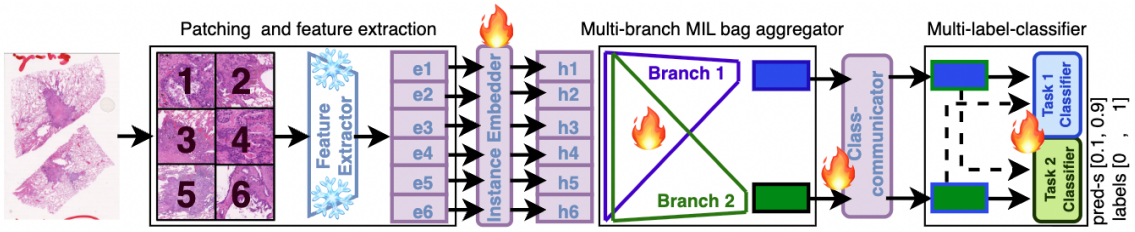

Figure 5.1: **The proposed class-dependency modelling framework.** Frozen feature extractor (patches 1-6 → embedding vectors e1-e6) outputs go into instance embedder (embedding vectors e1-e6 → hidden vectors h1-h6), which enter multi-branch MIL pipeline, then the class-communicator module, which reweighs every bag embedding from each branch using bag embeddings from other branches. Updated embeddings pass to the multi-label classifier. Dashed lines show that the classifier can also pass information between different tasks. Linear classifier achieves it by sharing the same weights for all tasks while communicating convolutional classifier accepts all class embeddings as input. A snowflake represents that the feature extractor is frozen, while fire represents the trainable modules.

previous works that used separate label-specific MIL aggregator branches. I used the feature extractor pre-trained on TCGA-lung data by Li et al. [117]. I concluded that modelling class dependencies enabled accurate joint prediction of lung cancer subtypes and patterns.

5.3.2 Contributions: Foundation Models and Mixed Supervision

I evaluated the best Dependency-MIL aggregator model described above on slides from four hospitals participating in the DART lung health programme [81]. These slides were collected later and were not used for training Dependency-MIL. I observed a drop in performance compared to the original test set. I decided not to focus on the domain generalisation methods in this work, so I split the DART dataset between training and evaluation sets. This made the distributions of the training and evaluation sets more similar, allowing me to focus on other modelling decisions without worrying about the distribution shift between the train and evaluation sets. After enriching the dataset with DART data, I went to evaluate the features extracted using recently released patch-level (UNI [46], Prov-GigaPath [195], Virchow-v1 [185], Phikon-v2 [74]) and slide-level (Prov-GigaPath [195], PRISM [160]) pathology foundation models concluding that my original feature

extractor, a ResNet18 pre-trained on TCGA-lung [117] was superior on my dataset. Finally, I used mixed supervision in the presence of partial labels by using non-exhaustive selections of diagnostic and benign regions available for OUH and DART sites. The results suggest that adding minimal extra time per subtyping annotation can improve classification performance, opening new avenues for low-resource annotation settings.

5.4 Dataset

To construct the dataset, I combined slides from three public lung cancer datasets: TCGA, TCIA-CPTAC [10], and Dartmouth Hitchcock Medical centre (DHMC) [190]. I also collected slides from Oxford University Hospitals (OUH) and four DART lung health programme sites. I constructed the dataset to be used for a multi-label classification task, where each task is a binary classification of whether the cancer subtype (LUAD, LUSC, Benign) or LUAD pattern (acinar, lepidic, micropapillary, papillary, solid) is *present* or *absent* on the slide. I also recorded the situations when the label was *unknown*.

TCGA lung cohort contains 541 digitised slides from 478 LUAD cases and 512 digitised slides from 478 LUSC cases. They are annotated as either LUAD or LUSC. Hence, I set the labels for all LUAD patterns as unknown on LUAD slides and absent on LUSC slides.

TCIA-CPTAC slides come from patients with either LUAD (591 slides) or LUSC (519 slides). Many of these patients also have slides with Benign Tissue (706 slides). For all LUSC and Benign slides, I marked LUAD patterns as absent. For LUAD slides, I created a mapping[1] to parse all unique case descriptions from the cohort information document. I marked the pattern as present if it was mentioned as present; otherwise - as unknown.

DHMC cohort contains 143 LUAD slides with predominant pattern labels. The patient identifiers are unavailable, so I assumed all slides to come from distinct patients. I marked all predominant patterns as present and all the other patterns as unknown.

[1] `https://github.com/GeorgeBatch/dependency-mil/blob/main/labels/source_copies_for_label_files/tcia_cptac_string_2_ouh_labels.csv`

OUH and DART lung cancer datasets included 164 slides from OUH and 193 slides from four DART sites. On LUAD slides, two expert thoracic pathologists explicitly annotated the LUAD patterns as present or absent. This is important since secondary LUAD patterns are significant for clinical prognosis and treatment but rarely reported in public datasets. None of the Benign and LUSC slides had any LUAD patterns. The annotation protocol and the data distribution are described in detail in Sections 3.2 and 3.2.3.

5.5 Modelling Class Dependencies

My goal was to train multi-label classification models to simultaneously classify lung cancer tissue into LUAD, LUSC, and benign tissue and identify the presence of adenocarcinoma patterns. My network, as shown in Figure 5.1, consists of (1) a frozen pre-trained feature extractor, (2) an instance embedder, (3) a multi-branch MIL bag aggregator, (4) a class-communicator module, and (5) a multi-label classifier module. Let us denote the whole slide dataset as $\{\mathcal{S}_k^K\}$ with K - the number of slides; $\mathcal{S}_k = \{x_n^{N_k}\}$ - a slide consisting of a set of patches with N_k - number of patches in the k-th slide and x_n is the n-th patch in the slide.

5.5.1 Feature Extraction

I followed a standard feature extraction process using the TIAToolbox library [145], which I extended to be able to work with OUH and DART slides scanned with a Roche Ventana scanner [2]. First, I created tissue masks using Otsu thresholding. Then I extracted patches of size 224x224 from whole slide images (WSIs) at 1 micron per pixel (10x magnification of most scanners) with no overlap filtering out the patches that had less than 10% of background tissue following the pre-processing of Xu et al. [195]. I utilised a ResNet18 [86] feature extractor pre-trained by Li et al. [117] on the TCGA lung cohort using SimCLR [49] to extract a $F_{extract} = 512$-dimensional feature embed-

[2] https://github.com/TissueImageAnalytics/tiatoolbox/pull/807

ding $e_n = f_{extract}(x_n), e_n \in \mathbb{R}^{F_{extract} \times 1}$ for each patch keeping the weights of $f_{extract}$ frozen and normalising with ImageNet constants. I chose the resolution of 1 μpp because Li et al. [117] pre-trained the ResNet18 feature extractor on patches from TCGA-lung cohorts extracted at 1 μpp.

5.5.2 Instance Embedder

I added an instance embedder ($f_{inst-embed}$) between the frozen feature extractor ($f_{extract}$) and the trainable multi-branch MIL bag aggregator to enable the model to adjust patch features before aggregating them. I used a single trainable linear layer, but other options like identity operator, adaptive average pooling, or multiple linear layers with non-linear activations can be explored thanks to the modular structure of the combined model. The instance embedder outputs features $h_n = f_{inst-embed}(e_n), h_n \in \mathbb{R}^{F_{inst-embed} \times 1}$.

5.5.3 Multi-branch MIL Bag Aggregator

The focus of this work was to explicitly model subtype dependencies. To investigate the ability of my proposed modules for multi-label weakly-supervised whole slide image classification, I used two popular MIL bag aggregators: AB-MIL [102] and DSMIL [117]. Both networks accept a bag of features with shape $N_k \times F_{inst-embed}$. I deploy a separate aggregation branch for each class and produce a bag embedding matrix $B = f_{mil}(\{h_1, ..., h_{N_k}\}), B \in \mathbb{R}^{C \times F_{mil}}$, where F_{mil} is the size of embedding corresponding to each binary class and C is the number of binary classes. In the baseline setting, a classification head is applied to the bag embedding, where each entry of the embedding matrix is scored, and a prediction for the corresponding binary class is obtained.

5.5.4 Class Communicator

I aimed to obtain a more accurate and holistic multi-label prediction by modelling the dependencies between the binary classes. Hence, I passed the bag embedding matrix

$B \in \mathbb{R}^{C \times F_{mil}}$ through a module that allows communications between the class entries B_i in the bag embedding matrix, and therefore, the following classification head can operate on an updated entry conditioned additionally on the entries of other classes $B_j, j \neq i$ (referred to as the context embedding in the NLP literature). For this purpose, I adopt two classic architectures, the original self-attention mechanism introduced by Bahdanau et al. [25] and the transformer self-attention mechanism by Vaswani et al. [184] with a single attention head. Let us denote the transformed bag embedding matrix as $\hat{B} = f_{cc}(B), \hat{B} \in \mathbb{R}^{C \times F_{mil}}$. In my experiments, I compared the two attention methods and the baseline where an identity operation was used ($B = \hat{B}$).

1) Bahdanau self-attention on input matrix $B \in \mathbb{R}^{C \times F_{mil}}$ operates as follows. First, **alignment scores** are computed for class embedding B_i and context embedding B_j:

$$A_{ij} = \mathbf{v}^T \cdot \tanh(\mathbf{W_1} \cdot B_i + \mathbf{W_2} \cdot B_j) \tag{5.1}$$

where $\mathbf{v}$ is a learnable parameter vector and $\mathbf{W_1}, \mathbf{W_2}$ are a learnable weight matrices. Then the **transformed class embeddings** $\hat{B}_i, i \in 1, ..., C$ are computed as weighted sums of all class embeddings, with the softmax-normalized alignment scores as weights:

$$\hat{B}_i = \sum_j \text{softmax}(A)_{ij} \cdot B_j \tag{5.2}$$

2) Transformer multi-head self-attention on an input matrix $B \in \mathbb{R}^{C \times F_{mil}}$operates as follows. First, **queries, keys, and values** are computed $Q = BW_Q$, $K = BW_K$, $V = BW_V$, where W_Q, W_K, W_V - learnable matrices. Second, **attention scores** A are computed for all pairs of classes:

$$A_{ij} = \frac{Q_i \cdot K_j^T}{\sqrt{d_k}} = \frac{Q_i \cdot K_j^T}{\sqrt{F_{mil}}}, \tag{5.3}$$

where d_k - dimension of the keys is equal to $F_{mil} = F_{mil}/1$ (the number of self-attention

heads is 1). Finally, **transformed class embeddings** $\hat{B}_i$ are computed as weighted sums of all class value vectors V_j, with the softmax-normalized attention scores as weights

$$\hat{B}_i = \sum_{j=1}^{C} \text{softmax}(A)_{ij} \cdot V_j \tag{5.4}$$

5.5.5 Multi-label Classifier

The classification head takes the transformed bag embedding matrix $\hat{B} \in \mathbb{R}^{C \times F_{mil}}$ as input and produces prediction $Y = f_{mc}(\hat{B}), Y \in \mathbb{R}^{C \times 1}$. In my experiments, I compared three alternative options for the multi-label classifier: a communicating convolutional layer, a depthwise-separable convolutional layer, and a linear layer.

1) Communicating convolutional layer. Prediction Y_c for each of the C classes is computed by applying C distinct convolutional filters $W_c \in \mathbb{R}^{C \times F_{mil}}$ to $\hat{B}$. This layer is implemented using a 1D convolution with kernel size F_{mil} (length of an embedding) and the number of input and output channels equal to C (number of classes) represented by the weights matrix W of shape $C \times C \times F_{mil}$ and C scalar biases.

$$Y_c = \sum_{j=1}^{F_{mil}} \sum_{k=1}^{C} W_{c,k,j} \cdot \hat{B}_{k,j} + b_c \tag{5.5}$$

where $W_{c,k,j}$ represents the weight of the convolutional filter corresponding to class c, applied to the feature j of class k in $\hat{B}$, and b_c is the bias term for class c.

2) Depthwise-separable convolutional layer.

$$Y_c = \sum_{j=1}^{F_{mil}} W_{cj} \cdot \hat{B}_{cj} + b_c \tag{5.6}$$

Each class prediction is computed using its own embedding $\hat{B}_c$. For each class, a filter of shape $1 \times F_{mil}$ is learnt. The layer consists of a $C \times F_{mil}$ matrix and C scalar biases.

3) Linear layer. A linear layer, containing weights of shape $F_{mil} \times 1$ and a scalar bias, is

shared among all classes. The prediction for each class is computed as:

$$Y_c = \sum_{j=1}^{F_{mil}} W_j \cdot \hat{B}_{c,j} + b \tag{5.7}$$

where W_j is the shared weight applied to the feature j, $\hat{B}_{c,j}$ is the corresponding input feature for class c, and b is the shared bias term.

5.5.6 Dataset Train/Test Split

I conducted the experiments on modelling class dependencies before receiving slides and annotations from sites participating in the DART lung health programme. Table 5.1 shows the distribution of known labels in the combined dataset and how I split the datasets into train-validation and test sets for the Dependency-MIL work.

Since I was planning to use the extractor pre-trained by Li et al. [117] on the **TCGA lung** dataset, I used the same train/test split as Li et al. [117] for TCGA lung slides to contain the data leakage to the training set and excluded the same 10 slides.

TCIA-CPTAC was the only part of my dataset that contained a considerable number of samples with benign tissue. Hence, I split its patients equally into train-validation and test sets stratifying on the LUAD vs LUSC label.

Since it is the first time the **DHMC** dataset is used for pattern presence prediction and not predominant pattern classification, I entirely put it into the test set to enable others to compare their algorithms with ours. Furthermore, this dataset comprises only LUAD slides, so if used for training, it would be enough for a classifier to learn how to distinguish the DHMC slides and predict LUAD for all of them in order to achieve a perfect score.

OUH was the only dataset with no unknown labels, so I put it entirely in the training set to give my models the best chance to learn. Moreover, using only public datasets for the test set enables other researchers to compare their results with mine.

	LUAD	LUSC	Benign	acinar	lepidic	micropapillary	papillary	solid
TCGA	428 / 106	403 / 109	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0
CPTAC	290 / 301	283 / 236	352 / 354	91 / 101	7 / 5	8 / 10	36 /4 1	29 / 13
DHMC	0 / 143	0 / 0	0 / 0	0 / 59	0 / 19	0 / 9	0 / 5	0 / 51
OUH	137 / 0	24 / 0	4 / 0	130 / 0	122 / 0	97 / 0	38 / 0	39 / 0
Total	855 / 550	709 / 345	356 / 354	221 / 160	129 / 24	105 / 19	74 / 46	68 / 64

Table 5.1: **Train-validation (1920 slides) / Test (1249 slides).** Columns 1-3: Data Distribution of adenocarcinoma (LUAD), squamous cell (LUSC), and normal/benign slides. Columns 4-8: Presence of five main LUAD patterns on the LUAD slides. The DHMC dataset is fully put into the test set. The OUH dataset is fully in the train set.

5.5.7 Masked Binary Cross-Entropy Loss

I used a masked binary cross-entropy loss to ignore predictions with unknown labels. Let N be the batch size and C be the number of classes, then:

- $X \in \mathbb{R}^{N \times C}$ are the logits (unnormalised predictions).
- $Y \in \{0, 1\}^{N \times C}$ are the binary target labels.
- $W \in \{0, 1\}^{N \times C}$ is the weight matrix with entry 1 for known and 0 for unknown labels. W acts as a mask to ignore predictions with unknown labels.

The binary cross-entropy loss for element n of class c element is given by:

$$\mathcal{L}_{n,c} = -\left(Y_{n,c} \log \sigma(X_{n,c}) + (1 - Y_{n,c}) \log(1 - \sigma(X_{n,c}))\right), \tag{5.8}$$

where $\sigma(z)$ is the sigmoid function.

Applying the elementwise weight mask W, the masked loss is:

$$\mathcal{L}_{n,c}^{\text{masked}} = W_{n,c} \cdot \mathcal{L}_{n,c}. \tag{5.9}$$

Summing over all batch elements and normalising by the sum of the weights gives:

$$\mathcal{L} = \frac{\sum_{n=1}^{N} \sum_{c=1}^{C} \mathcal{L}_{n,c}^{\text{masked}}}{\sum_{n=1}^{N} \sum_{c=1}^{C} W_{n,c}} = \frac{\sum_{n=1}^{N} \sum_{c=1}^{C} W_{n,c} \cdot \mathcal{L}_{n,c}}{\sum_{n=1}^{N} \sum_{c=1}^{C} W_{n,c}}. \tag{5.10}$$

This weights matrix definition and the normalization by the sum of weights ensures that every known label is weighted equally and samples with more known labels are weighted higher in the batch loss calculation.

5.5.8 Implementation Details

For all experiments, I used the same train-validation-test setup without resampling the data. I used features computed by a frozen feature extractor without data augmentation.

Training. The models were trained for 10 epochs with a batch size of 1 using Adam optimiser [110] with a learning rate of 1e-4 for AB-MIL and 2e-4 for DSMIL, weight decay 0.005, beta's (0.5, 0.9). I used a Cosine Annealing [123] scheduler to vary the learning rate with T_{max}=num epochs and η_{min}=5e-6. During training, the parameters of $f_{inst-embed}$, f_{mil}, f_{cc}, and f_{mc} were optimized (see Sections 5.5.2-5.5.5).

Validation. Working in a multi-label setting in the presence of unknown labels, I computed the metrics separately for each class, only considering samples with known labels. I computed AUC, Precision, Recall, and F1 scores. Then, I chose the threshold that would result in the best possible accuracy on the whole validation set and binarised my predictions. Using the binarised predictions and the true labels, I computed accuracy for each class.

Finally, I computed the subset accuracy for all classes together, which is defined as the proportion of slides for which the predictions were entirely correct. The predictions for a slide are considered entirely correct if and only if all known labels are predicted correctly. If we predict 8 binary labels (LUAD, LUSC, Benign, and 5 LUAD patterns) for a slide with all LUAD pattern labels unknown, it would not matter what we predicted for the LUAD patterns, even if, in reality, these predictions were incorrect.

Following Li et al. [117], I recorded the weights and the best-case validation thresholds if the score S_{new} combined from the subset accuracy Acc_{subset} and per-class AUC scores,

Architecture	Acc.	LUAD	LUSC	Benign	acinar	lepidic	m-papillary	papillary	solid
AB-MIL + identity + linear	13.69	81.48	64.31	45.52	76.39	53.32	73.30	81.16	67.03
AB-MIL + transformer + linear	17.61	84.49	26.50	31.28	86.72	88.11	98.04	83.65	87.95
AB-MIL + transformer + c_conv	82.63	95.06	96.16	99.70	96.58	98.59	99.95	98.81	95.28
AB-MIL + transformer + ds_conv	83.19	94.90	96.06	99.73	95.99	98.73	99.93	98.34	95.36
DS-MIL + identity + linear	30.26	83.50	82.71	67.15	80.72	81.80	96.98	79.88	85.37
DS-MIL + transformer + linear	32.83	84.69	85.99	78.55	94.85	80.27	98.72	80.55	88.03
DS-MIL + transformer + c_conv	69.58	88.75	94.68	97.85	92.60	92.13	98.44	86.66	89.95
DS-MIL + transformer + ds_conv	84.39	93.77	95.85	99.30	95.51	98.70	99.91	99.00	94.68

Table 5.2: **Test performance summary on comb-8 data. Column 2: Subset accuracy** calculated as the proportion of samples with fully correct predictions for all considered labels. **Columns 3-10: ROC AUC** are calculated separately for each task on the test set. Predictions with unknown labels are ignored for subset accuracy and ROC AUC calculations. Proportions of test set samples with known labels: LUAD 1, LUSC 1, Benign 1, acinar 0.68, lepidic 0.58, micropapillary 0.57, papillary 0.60, solid 0.61.

improved compared to the previous best combined score S_{best}.

$$S_{\text{new}} = \frac{Acc_{\text{subset}} + \sum_{i=1}^{C} \text{AUC}_i}{1 + C} \tag{5.11}$$

If if $S_{\text{new}} > S_{\text{best}}$, I saved weights, thresholds, and updated S_{best}.

Testing. Using the weights that resulted in the best score on the validation set, I would compute the test metrics using the saved binarisation thresholds instead of computing new ones on the test set.

5.5.9 Results and Discussion

I evaluated the proposed modules using two popular weakly-supervised WSI classification backbones: AB-MIL [102] and DSMIL [117]. I performed ablation studies to demonstrate the performance gain of the proposed class communicator modules and multi-label classifier modules. Subset accuracy on the test set, defined as the proportion of samples for which all labels (including subtypes and patterns) were correctly predicted, are reported. The performance of individual tasks was quantified using the standard metric - area under the ROC curve (ROC AUC).

I tested the model in 4 scenarios: 1) comb-2: LUAD and LUSC; 2) comb-3: LUAD, LUSC, benign; 3) comb-8: LUAD, LUSC, benign, and 5 LUAD patterns; 4) comb-5:

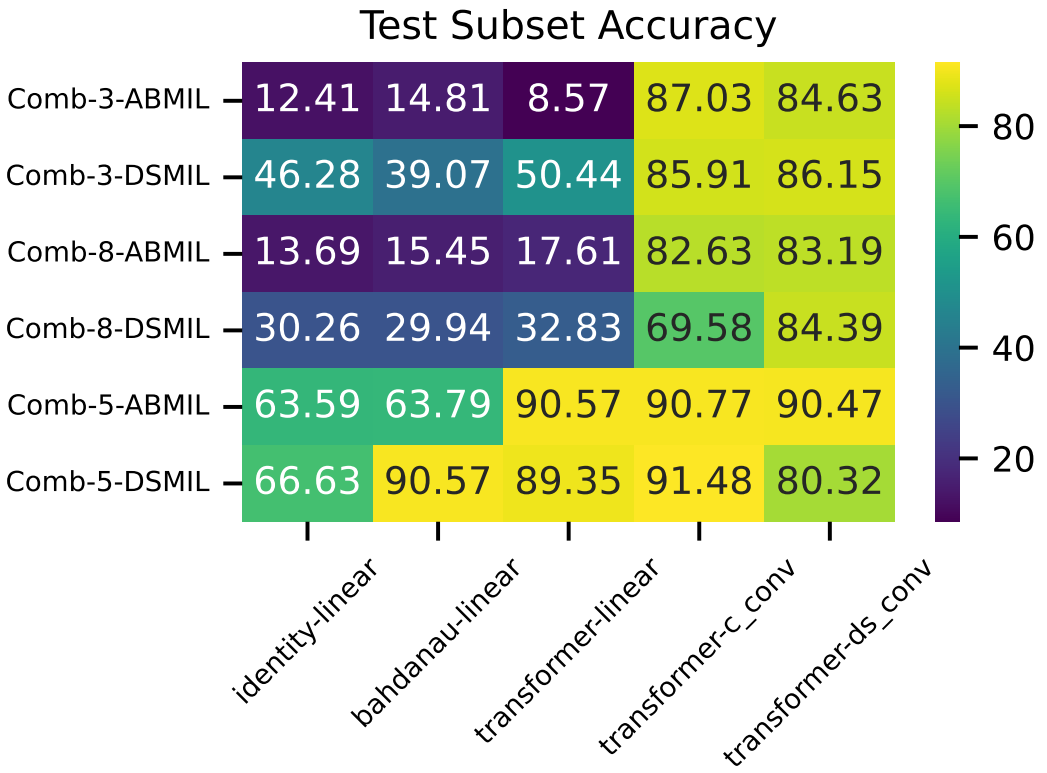

Figure 5.2: Subset accuracies of data-model pairs. Y-labels: tasks (comb -3, -5, -8) and the MIL-aggregator architecture (AB-MIL, DSMIL). X-labels: combinations of class-communicator (identity vs transformer) and multi-label classifier module (linear, communicating convolution - "c_conv", and depthwise-separable convolution - "ds_conv".)

only 5 LUAD patterns. I present the subset accuracies and ROC AUC scores for the complete multi-label classification task (comb-8) in Table 5.2. Figure 5.2 summarises the subset accuracies for the above scenarios.

The results demonstrate that the baseline weakly-supervised classification models can be deficient in multi-label settings and achieve unsatisfactory performance (subset accuracy $< 60\%$, first column of Figure 5.2). After applying the class communicator module, the classifier successfully learns to differentiate the 5 adenocarcinoma patterns (comb-5 in Figure 5.2) but still shows unsatisfactory performance for simultaneously predicting cancer subtypes and adenocarcinoma patterns (comb-8 in Figure 5.2). After incorporating the proposed multi-label classifier module, implemented via a communicating convolutional layer or depthwise-separable convolutional layer, a large performance gain is also

obtained in the 8-label task, where subtypes and patterns are classified together.

A plausible explanation is that the baseline models assume that learned bag embeddings for different classes are independent. Thus, each entry of a bag embedding encodes all information needed to infer the label for that class. The baseline classification head is also shared across the classes, which assumes that the classes are lateral (no hierarchical dependency). The proposed class communicator module, implemented via Bahdanau [25] or Transformer [184] attention, allows information to flow between bag embedding entries, enabling each class entry to incorporate information from other classes. This captures dependencies between classes and yields an enriched and more robust bag embedding. Finally, because the adenocarcinoma patterns are essentially sub-classes within LUAD, the underlying structure among the labels is not entirely lateral but somewhat hierarchical. The multi-label classifier module, which offers more flexibility than a standard linear classification head, can accommodate this complexity to a certain extent.

5.5.10 Initial Conclusions

I proposed a novel class-dependency modelling method that can be readily incorporated into weakly-supervised whole slide classification models for multi-label problems. The model allows information about the presence of other classes in the sample to be shared between the classification branches. Incorporating the class dependency into the model architecture resulted in a more accurate joint prediction of broad and fine lung cancer subtypes. In addition, I present a new dataset combining UK lung screening (DART) and public datasets to learn lung cancer classes and adenocarcinoma patterns jointly.

This work has two limitations: 1) less than 10% of LUAD slides in the test set have positive labels for lepidic or micropapillary pattern presence, and 2) most adenocarcinoma patterns reported in DHMC and TCIA-CPTAC datasets are predominant, which makes their identification easier than for secondary patterns.

5.6 DART Data, Foundation Models, Mixed Supervision

Unlike the experiments presented in Section 5.5, the experiments presented in this section were conducted after receiving slides and annotations from four sites participating in the DART lung health programme in January 2025.

5.6.1 Evaluation on the DART datasets

I evaluated the best 8-label binary classification model (DS-MIL aggregator, transformer class communicator, depthwise separable convolution classifier) on DART slides. The label's distribution of the DART dataset can be found in Section 3.2.3. Table 5.3 shows the evaluation results on the DART dataset. We observe that the performance is worse than on the test dataset considered in Section 5.5. Where the subset accuracy on the test set was 84.39%, the subset accuracy achieved on the combined DART dataset was 28.5%, with the model showing the highest subset accuracy of 39.06% for DART site 3. The AUC scores also dropped from being all above 90% to under 90% for all 5 LUAD pattern labels (out of 8 classes) on the combined DART dataset.

The drop in performance can be explained by the distribution shift between the training data and the DART data and the limitations of the test set. The LUAD pattern labels in the test set were gathered from TCIA-CPTAC and DHMC datasets, which meant that the majority of these labels referred to predominant LUAD patterns, making their identification easier than the secondary LUAD patterns available for the DART dataset.

	Acc.	LUAD	LUSC	Benign	acinar	lepidic	micropapillary	papillary	solid
DART 1	16.67	85.71	98.08	86.42	73.68	76.56	72.00	75.00	68.75
DART 2	29.79	97.44	97.62	88.01	81.61	77.90	86.88	91.28	76.67
DART 3	39.06	98.91	98.56	97.79	88.99	90.49	92.46	78.38	81.89
DART 4	21.15	88.54	93.61	86.03	90.18	76.63	89.65	82.00	66.83
Combined	28.50	94.03	96.27	90.09	84.15	79.11	86.73	79.98	71.02

Table 5.3: **Evaluation Performance on the DART datasets. Column 2: Subset accuracy** calculated as the proportion of samples with fully correct predictions for all considered labels. **Columns 3-10: ROC AUC** are calculated separately for each task. Our pathologists annotated this dataset, so all labels are known.

	LUAD	LUSC	Benign	acinar	lepidic	micropapillary	papillary	solid
TCGA	426 / 105	403 / 109	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0
CPTAC	295 / 296	265 / 254	354 / 352	97 / 95	9 / 3	4 / 14	47 / 30	12 / 30
DHMC	0 / 143	0 / 0	0 / 0	0 / 59	0 / 19	0 / 9	0 / 5	0 / 51
OUH	112 / 25	18 / 5	3 / 1	106 / 24	97 / 25	76 / 21	30 / 8	33 / 6
DART 1	18 / 5	2 / 2	2 / 1	14 / 5	10 / 1	9 / 1	5 / 1	5 / 3
DART 2	19 / 7	11 / 1	5 / 5	12 / 6	6 / 5	14 / 1	4 / 0	8 / 4
DART 3	30 / 6	10 / 2	14 / 2	22 / 4	21 / 1	26 / 2	8 / 1	11 / 1
DART 4	24 / 12	4 / 1	11 / 0	20 / 6	13 / 5	15 / 2	1 / 1	13 / 4
Total	924 / 599	713 / 374	388 / 361	271 / 202	156 / 64	144 / 59	95 / 46	82 / 100

Table 5.4: **Train-validation (2025 slides) / Test (1334 slides).** Columns 1-3: Data Distribution of adenocarcinoma (LUAD), normal/benign, and squamous cell (LUSC) slides. Columns 4-8: Presence of five main adenocarcinoma patterns on the adenocarcinoma slides. The DHMC dataset is fully added into the test set. The OUH and DART datasets are stratified by the cancer class and split in an 80/20 ratio.

5.6.2 Using Patch- and Slide-level Pathology Foundation Models

Instead of focusing on developing methods for domain generalisation, I decided to employ foundation models released in abundance in 2024. The evaluation on the dataset combined from TCGA, TCIA-CPTAC, DHMC, OUH, and DART. Table 5.4 shows the new data distribution of train-validation and test sets. I kept the split of the TCGA and TCIA-CPTAC datasets described in Section 5.5. I also kept the DHMC fully in the test set because all DHMC slides come from adenocarcinoma tumours. Furthermore, only predominant pattern presence is reported, so it would be enough for the model to learn to recognise the domain of DHMC and predict LUAD and all LUAD patterns as present all the time (1 pattern present, other patterns unknown), getting a perfect score. I split both OUH and DART datasets in an 80/20 ratio into train-validation and test splits, stratifying on the LUAD and LUSC labels and ensuring that slides from one patient appear either in the train-validation or in the test sets.

I showed in Chapter 4 that until a better pretext task is found to pre-select a promising foundation model, one needs to evaluate multiple available foundation models on a new dataset. So, in addition to ResNet18 pre-trained by [117], I chose UNI [46], Prov-GigaPath [195], Phikon-v2 [74], and Virchow-v1 [185] patch level foundation models as my feature extractors. I also considered Prov-GigaPath and PRISM slide feature extrac-

Architecture	Acc.	LUAD	LUSC	Benign	acinar	lepidic	m-papillary	papillary	solid
ResNet18 (0.5 mpp) + AB-MIL	**73.91**	**76.79**	**95.87**	**97.43**	70.02	81.07	83.52	**89.95**	60.03
Virchow-v1 + AB-MIL	40.41	58.20	89.84	83.67	63.89	68.05	74.67	77.99	36.52
Phikon-v2 + AB-MIL	61.32	71.48	92.83	84.05	**77.86**	**93.86**	**90.59**	86.49	74.33
UNI + AB-MIL	44.45	58.03	89.24	85.47	63.02	66.15	73.34	73.77	41.77
Prov-GigaPath + AB-MIL	29.69	58.08	84.13	74.86	72.36	83.93	72.00	64.99	**80.17**
ResNet18 (0.5 mpp) + DS-MIL	**69.04**	**71.45**	**97.55**	**98.84**	68.96	69.76	77.63	**86.97**	42.10
Virchow-v1 + DS-MIL	27.47	62.55	89.94	87.43	65.23	67.45	77.87	79.80	39.75
Phikon-v2 + DS-MIL	59.15	69.62	93.83	95.52	68.29	75.99	76.91	84.22	45.05
UNI + DS-MIL	46.93	62.93	90.20	86.31	65.61	66.67	74.82	80.33	43.68
Prov-GigaPath + DS-MIL	34.03	55.48	83.20	75.95	**70.94**	**82.64**	**78.69**	77.28	**56.58**
PRISM Slide	38.98	59.19	**85.93**	79.89	63.05	72.41	72.76	70.92	51.43
Prov-GigaPath Slide	**40.18**	**67.22**	83.06	**84.03**	**79.58**	**90.95**	**85.39**	**71.12**	**76.58**

Table 5.5: **Test set performance summary on comb-8 data. Column 2: Subset accuracy** calculated as the proportion of samples with fully correct predictions for all considered labels. **Columns 3-10: ROC AUC** are calculated separately for each task on the test set. Predictions with unknown labels are ignored for subset accuracy and ROC AUC calculations. Except for ResNet18 (1 mpp), all other models used patches extracted at 0.5 mpp. The last two rows represent results achieved by fitting a 2-layer network on top of PRISM and Prov-GigaPath slide-level features.

tors pre-trained on top of Prov-GigaPath and Virchow-v1 patch-level embeddings, respectively. I followed the feature extraction procedure described in Section 5.5.1 extracting patches at 0.5 μpp and normalising with ImageNet constants for all 4 foundation models following the recommendations of their authors. In order to use the TIAToolbox [145] functionality, I contributed to it by integrating the foundation models into the TIAToolbox package [3].

When training models on a new dataset on top of frozen patch embeddings, I always used linear instance embedder, transformer class communicator, and depthwise separable convolution classifier - the combination shown to perform best in Section 5.5.9. I experimented with both AB-MIL and DS-MIL aggregators. The models did not converge on a new dataset with training hyperparameter settings described in Section 5.5.8. Inspired by the strong benchmarking results reported by Filiot et al. [74], I removed the learning rate scheduler and changed the Adam optimiser betas from (0.9, 0.5) to (0.9, 0.999), which enabled the training of the aggregator model to converge when using the ResNet18 features [117]. Other aggregator models could not escape the local minima (underfitting), so I also experimented with setting all hyperparameters to ones described by Filiot et al. [74],

[3] `https://github.com/TissueImageAnalytics/tiatoolbox/pull/856`

increasing the batch size from 1 to 16, the learning rate from 0.0001 to 0.001, randomly subsampling 5000 patches from a slide at each iteration, and training for 50 epochs. These changes made it possible to achieve the results presented in Table 5.5. I observed that the ResNet18 [117] feature extractor was superior when used with the AB-MIL aggregator, followed by Phikon-v2 [74]. With the DS-MIL aggregator, the ResNet18 [117] was still superior when considering subset accuracy and cancer subtyping (LUAD, LUSC, Benign) but performed worse than Prov-GigaPath on 4 out of 5 LUAD pattern prediction tasks.

The general superior performance of ResNet18 can be partially explained by data leakage that occurred during training on the TCGA lung cohort performed by Li et al. [117]. However, I contained data leakage to the training set by using the same train/test split of TCGA slides as Li et al. [117]. Furthermore, TCGA samples accounted for only 214 out of 1334 testing samples. The fact that a smaller feature extractor pre-trained only on lung cancer images can be competitive for lung cancer classification against much larger foundation models pre-trained on the pan-cancer datasets with more than 10x slides suggests that specialisation (focusing on 1 cancer type) can be more effective than generalisation (learning features for all cancer types) if one has access to a training cohort for specialisation. Alternatively, foundation models can have more random parameters, making overfitting likely in data-sparse scenarios, especially when used without augmentations like patch subsampling or train-time stain augmentation.

A simple neural network consisting of 2 linear layers with 128 hidden and 8 output dimensions showed better results than a single linear layer with 8 output dimensions when trained on top of the frozen slide-level features extracted with PRISM and Prov-GigaPath slide-level embedders. It showed the best performance when trained with a batch size of 16 for 100 epochs with Adam Optimiser, which had the learning rate set to 0.001 and betas (0.9, 0.999). Prov-GigaPath slide-level embeddings were superior to PRISM embeddings. Using both PRISM and Prov-GigaPath slide embeddings, followed by a linear or a 2-layer classifier, performs better than their patch-level counterparts (Virchow-v1

and Prov-GigaPath), which aligns with results presented by Shao et al. [161], who show that fine-tuning pre-trained slide-level foundation models works better than starting from pre-trained patch-level models and training slide-level models from scratch. However, my results disagree with findings from Neidlinger et al. [137], who found that "the original tile embeddings consistently outperformed their slide-level counterparts and the performance of the encoded tile embeddings was driven by the quality of the original tile embeddings and not by the slide encoder".

5.6.3 Mixed Supervision: Focusing on Diagnostic Regions

After determining that ResNet18 [117] was generally better than other feature extractors, I decided to use the region annotation provided by our pathologists for mixed slide- and patch-level supervision in order to focus the learning on the diagnostic regions, as shown in Figures 5.3 and 5.4. Pathologists were asked to select diagnostic regions they used for

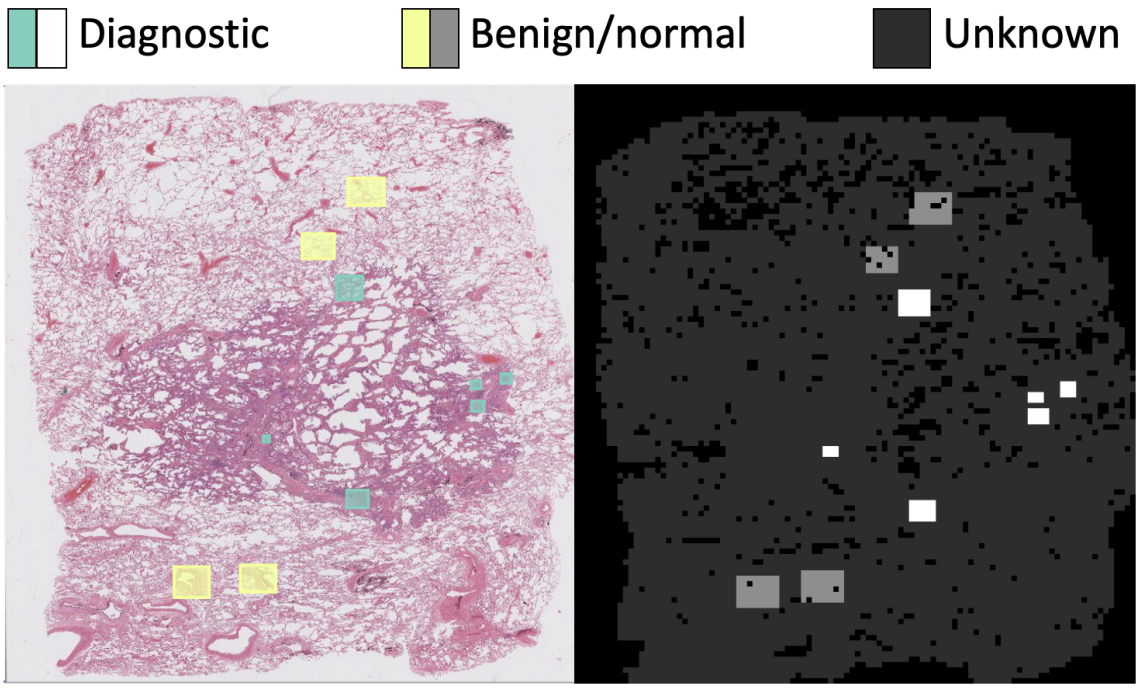

Figure 5.3: **Left: region annotation of a whole slide image.** Green rectangles represent diagnostic tissue, yellow rectangles - benign tissue. **Right: patch-level status.** White represents diagnostic patches, light grey - benign patches, dart grey - patches with unknown status, black - background patches.

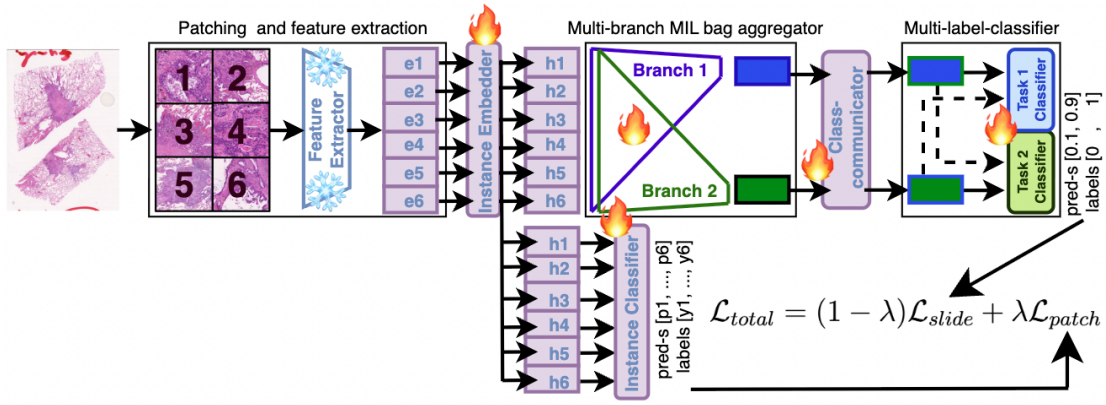

Figure 5.4: **The proposed class-dependency modelling framework with mixed supervision.** A snowflake represents that the feature extractor is frozen, while fire represents the trainable modules. **Top:** Frozen feature extractor (patches 1-6 → embedding vectors e1-e6) outputs go into instance embedder (embedding vectors e1-e6 → hidden vectors h1-h6), which enter multi-branch MIL pipeline, then the class-communicator module, which reweighs every bag embedding from each branch using bag embeddings from other branches. Updated embeddings pass to the multi-label classifier. Dashed lines show that the classifier can also pass information between different tasks. A linear classifier achieves it by sharing the same weights for all tasks, while communicating convolutional classifier accepts all class embeddings as input. **Bottom:** Patch-level supervision is achieved by adding an instance classifier that takes the outputs of the instance embedder (h1-h6) and outputs the probabilities for each patch to be diagnostic or benign.

subtyping slides from the OUH and DART datasets, and some regions with non-diagnostic (benign) tissue. Section 3.2 provides a detailed description of the annotation protocol.

Figure 5.3 demonstrates how an annotation from the pathologist of a resection slide containing tumour tissue from OUH is converted into a patch-level tissue mask. If a slide is subtyped as "Benign", it does not need to have the region annotations since all its tissue patches can be considered benign, so all benign slides from OUH, DART, and TCIA-CPTAC datasets received patch-level benign labels. Slides containing tumours from TCGA, TCIA-CPTAC, and DHMC did not have patch-level labels.

I constructed patch-level labels and weights analogously to slide-level labels and weights for each patch. Patches with known labels received a weight of 1, while patches with unknown labels received a weight of 0. Patch labels were set to 1 for diagnostic and 0 for benign regions. As shown in Figure 5.4, I fitted a linear classifier layer $f_{inst-class}$ on top of the instance embedded $f_{inst-embed}$ (Section 5.5.2) and used the masked binary

cross-entropy loss (Section 5.5.7) to optimise $f_{inst-embed}$ and $f_{inst-class}$.

Recall the notation where K is the number of slides, each slide $\{S_k\}, k \in \{1, ..., K\}$ consists of N_k patches, and C is the number of binary slide-level classification tasks. Adding $f_{inst-class}$ and instance classification masked binary cross-entropy loss resulted in mixed supervision with the total loss for slide S_k:

$$\mathcal{L}_{total} = (1-\lambda)\mathcal{L}_{slide} + \lambda\mathcal{L}_{patch} = (1-\lambda)\frac{\sum_{c=1}^{C} W_c \cdot \mathcal{L}_c}{\sum_{c=1}^{C} W_c} + \lambda\frac{\sum_{n=1}^{N_k} W_n \cdot \mathcal{L}_n}{\sum_{n=1}^{N_k} W_n}, \quad (5.12)$$

where λ is the weight parameter that balances the contribution of slide-level and patch-level losses into the total loss. As a result, instance embedded module $f_{inst-embed}$ is optimised with both slide and patch-level losses, instance classifier module $f_{inst-class}$ - only with the patch-level loss, while MIL aggregator f_{mil}, class communicator f_{cc} and multi-label classifier f_{mc} modules - only with slide-level loss.

This resulting mixed supervision forces the instance embedded module to learn a better patch embedding transformation as demonstrated by the improved classification performance results presented in Table 5.6. I kept all training settings as in Section 5.6.2 and experimented with λ values of (0.25, 0.5, 0.75), achieving the best results with $\lambda = 0.25$.

Architecture	Acc.	LUAD	LUSC	Benign	acinar	lepidic	m-papillary	papillary	solid
ResNet18 + AB-MIL	73.91	76.79	**95.87**	97.43	70.02	81.07	83.52	**89.95**	60.03
ResNet18 + AB-MIL + MS	**75.33**	**84.62**	93.96	**98.11**	**77.90**	**85.38**	**83.98**	87.60	**70.77**
ResNet18 + DS-MIL	69.04	71.45	**97.55**	**98.84**	68.96	69.76	77.63	86.97	42.10
ResNet18 + DS-MIL + MS	**72.19**	**86.93**	97.52	97.64	**85.40**	**88.26**	**90.38**	**92.43**	**79.64**

Table 5.6: **Test performance summary on comb-8 data. Column 2: Subset accuracy** calculated as the proportion of samples with fully correct predictions for all considered labels. **Columns 3-10: ROC AUC** are calculated separately for each task on the test set. Predictions with unknown labels are ignored for subset accuracy and AUC calculations. "MS" stands for mixed supervision, i.e., using slide and tile-level labels for training.

5.7 Conclusions

In this chapter, I present a new dataset combining TCGA, TCIA-CPTAC, and DHMC public lung cancer datasets with in-house datasets from OUH and DART to learn lung cancer classes and adenocarcinoma patterns jointly. For TCIA-CPTAC I manually parsed the slide text descriptions from the cohort information file and released the extracted structured labels. I proposed to use class communicator and classifier modules on top of multi-branch MIL aggregators in order to explicitly model class dependencies. These modules can be readily incorporated into weakly-supervised whole slide classification models for multi-label problems. The results obtained on the proposed dataset support the hypothesis that enabling the exchange of information between different class-specific branches increases the overall joint classification performance of lung cancer subtypes and adenocarcinoma patterns. This work was published at ISBI 2024 conference [33].

After receiving DART annotations in January 2025, I was able to evaluate the proposed models on a new dataset that contained higher-quality labels than present in the original test set, observing the drop in performance and concluding that (1) my original test set had limitations coming from high quantity of missing labels and having only predominant LUAD pattern labels for slides from TCIA-CPTAC and DHMC datasets and (2) the trained model was not yet robust enough to perform on unseen data distribution.

I then focused on evaluating the effect of using different foundation models on the new combined dataset, which included DART slides in both the training and test sets. I contributed to TIAToolbox, adding support for non-tiled TIFF files scanned with the Roche Ventana DP200 scanner and open-source foundation models.

Finally, I achieved results supporting the hypothesis that adding minimal extra annotation time by asking the pathologists to select diagnostic regions they use for subtyping as well as a few benign regions they can spot quickly can improve the classification performance by enabling mixed supervision in the presence of partial labels, i.e. using both weak slide-level labels and patch level labels when both can have missing labels.

Chapter 6

Conclusions and Future Directions

Contents

6.1 Summary of Contributions

By design, my doctorate project was part of a larger effort - the DART Lung Health Project [81]. The DART project aimed to improve everything connected to lung cancer diagnosis, including diagnostics from CT scans, pathology images, and blood biomarkers.

I aimed to (1) construct an annotated lung histology dataset and improve upon the state-of-the-art models for classifying lung cancers from pathology images and (2) combine my lung histology model with a CT model that was being developed by Dr Mengran Fan, my colleague in the research group and the DART work package, at the same time as I was working on my histology model. Due to the complications with acquiring the CT data described in Chapter 1, I focused my thesis entirely on aim 1 - classifying lung cancers from histopathology images.

To facilitate work on the project, I contributed to the development of the dataset in ways that were not research contributions in and of themselves but which were essential for my subsequent research and can be reused by other researchers working with the DART data. These contributions include:

- a multi-stage annotation protocol that included slide-subtyping, region selection, and region annotation;
- software for parsing AIDA [14] annotations to labels for machine learning models;
- an annotation time-sharing strategy implemented through interactive tracker sheets to enable three pathologists to annotate the same pool of histopathology slides using two annotation protocols (mine and of the Roche team) without annotating the same slide twice.

6.1.1 Active Data Enrichment

This section summarises the contributions described in Chapter 3. Parts of this chapter were published and presented at the MICCAI 2022 CaPTion Workshop [31] and at the 2022 Digital Pathology Workshop hosted by Roche in the Royal College of Pathologists.

Prioritising Pre-selected Regions, Ranking Curve AUC. After creating the detailed region annotation protocol aimed at a low-data setting (few slides), I explored supervised and unsupervised methods for optimizing the region annotation process described in Sec-

tion 3.4. I proposed a new metric, Ranking Curve AUC, to evaluate how well a method can prioritise the regions with a pattern of interest when the amount of annotation time is unknown. I showed that when presented with very few samples of a morphological pattern, one can effectively use unsupervised retrieval to rank diagnostic regions pre-selected by pathologists. I also showed that having only 20 regions with the pattern of interest present and 30 regions without the pattern of interest, it is possible to train a simple classifier to prioritise regions for annotation with the pattern present. Furthermore, annotating prioritised images and adding them to the training set results in a better test set performance than annotating random images. Both unsupervised and supervised methods depend on the availability of a suitable feature extractor. Although I achieved positive results in this chapter, a more thorough investigation of the long-term impact of the enrichment strategies on the annotation process would be beneficial. The proposed enrichment methods might lead to prioritizing images that fit the subtype/pattern distribution we already know instead of expanding our knowledge about these subtypes/patterns. While this can be a desired effect at the start of the annotation process since little is known of the data distribution, the proposed strategy might not be optimal if used long term. I could not validate this because many more slides became available while the pathologist annotation time remained scarce. Consequently, I had to abandon the detailed region annotation stage altogether.

Region Selection. I attempted to mine regions with patterns under-represented in the annotated dataset from slides. This was the first experiment I conducted in my doctoral project. No suitable feature extractors were available in the public domain when the experiments were conducted, and I used a truncated ResNet-50 [124] model that was pre-trained on ImageNet. None of the 20 regions I selected to enrich for the keratinisation pattern had it in reality. Although I got a negative result, other researchers have explored this avenue of research. In 2024 Qiu et al. [148] showed that it is possible to efficiently mine regions from slides that are similar to query (prototype) regions from a visual-language database using a visual-language foundation model like PLIP [97].

Slide Selection. Following the release of slide-level foundation models, Prov-GigaPath [195] and PRISM [160], I repeated my unsupervised retrieval experiments, but this time, I aimed to prioritise slides with Squamous Cell Carcinoma and Typical Carcinoid slide labels under-represented in my dataset, which was dominated by Adenocarcinomas. The results showed that both slide-level features and the text-based retrieval strategies performed better than random for enriching for the subtypes of interest.

6.1.2 Pathology Foundation Models

This section summarises the contributions described in Chapter 4. Parts of this chapter were published and presented at the MICCAI 2024 DEMI Workshop [34]. This paper got the Best Paper Award of the DEMI workshop.

LC25000 dataset [35] cleaning. I identified a data leakage problem in a popular tile-level lung and colon cancer dataset that has been cited in more than 440 works, according to Google Scholar (at the time of writing). I created a semi-automatic clustering pipeline that relied on UNI [46] to extract image features and curated the dataset, successfully producing LC25000-clean. Before UNI was released, I tried curating it using truncated ResNet50 pre-trained on ImageNet and ResNet18 pre-trained on TCGA lung patches [117]. However, neither model produced strong enough embeddings to separate the clusters of augmented images. If I were to do it before UNI was released, I would generate more augmentations on the LC25000 dataset and train a rotational-invariant feature extractor using a self-supervised learning method like SimCLR [49]. I believe that this dataset-specialised feature extractor could have been good enough for the task.

LC25000 dataset classification. I benchmarked different natural-image and pathology-specific feature extractors on lung and colon cancer classification tasks using the original LC25000 and curated LC25000-clean. Although there was a drop in classification performance when using the curated version of the dataset, strong feature extractors like UNI [46] and Prov-GigaPath [195] followed by a linear layer, were enough to solve the classi-

fication problem. Having achieved a >99% accuracy on a balanced (250 images per class) cleaned dataset, I conclude that the classification task on the LC25000 has been solved.

Clustering Augmented Patches as a pretext Task. I noticed a correspondence between the ability of the models to cluster augmented patches from the LC25000 dataset and the downstream classification performance. This sparked an idea to check if the relationship holds for clustering augmented patches extracted from whole slide images and downstream performance on slide-level classification tasks. This idea was received with a lot of interest at the 2024 MICCAI DEMI workshop, where I presented my paper [34]. I explored it by sampling, augmenting, and clustering patches from TCIA-CPTAC lung [10], CHAMELYON16 breast metastasis [122], TCGA lung, and combined Oxford University Hospitals (OUH) + DART lung datasets. For TCIA-CPTAC and CAMELYON16 datasets, I compared the clustering performance with the classification performances reported in benchmarking studies. For TCGA and OUH+DART datasets, I performed downstream classification myself. The results did not support the hypothesis that the clustering performance could be used to to choose a promising foundation model. Although I did not achieve positive results, I believe that this avenue of research is worth exploring since it has the potential to enable choosing a suitable foundation model at a fraction of the computational cost compared to extracting all patch embeddings with many different foundation models. It is possible that using better and stronger augmentations, e.g. stain augmentation, before performing the clustering can help to pick foundation models that are not just rotationally invariant, but that are good at extracting clinically-useful features.

6.1.3 Subtyping Lung Cancers

This section summarises the contributions described in Chapter 5. Parts of this chapter were published and presented at the ISBI 2024 [33] and the BTOG 2024 [32] conferences.

Dataset. I combined three public lung cancer datasets TCGA, TCIA-CPTAC [10], and DHMC [190] with the datasets collected from Oxford University Hospitals (OUH) and

DART sites to perform a multi-label classification of lung cancer subtypes (adenocarcinoma, squamous cell carcinoma, benign) and main adenocarcinoma patterns (acinar, lepidic, micropapillary, papillary, solid). For the TCIA-CPTAC dataset, I parsed the cohort information document to extract structured pattern labels from text descriptions.

Dependency-MIL. I proposed a modular method for modelling class dependencies and jointly learning the presence of lung cancer subtypes and patterns from weakly supervised slide-level labels. This method enabled popular MIL-based aggregators, AB-MIL [102], and DS-MIL [117], to learn robust bag embeddings by sharing information between the class branches. Class connector and multi-label classifier modules can be added to any multi-branch aggregator used for classification in a multi-label setting.

Evaluation of Foundation Models. I evaluated 5 pathology-specific feature extractors, including recently released foundation models on the proposed dataset. ResNet18 pre-trained on TCGA lung dataset [117] showed superior performance, suggesting that smaller models specialised in specific tissue types can result in better downstream performance than general foundation models. Furthermore, simple classifiers pre-trained on top of slide embeddings extracted with slide-level foundation models performed better than their patch-level counterparts contrary to the results reported by Neidlinger et al. [137].

Mixed Supervision with Partial Labels. Finally, I showed that asking pathologists to select regions they used for making subtyping annotations can be used for mixed supervision to improve classification performance. Unlike exhaustive tumour annotations that can increase the time a pathologist spends per slide tenfold (from 3 minutes to 30), marking regions already used for subtyping with rectangular bounding boxes takes virtually no extra time.

6.2 Directions of Future Work

This section outlines possible avenues for future work.

6.2.1 Domain Generalisation

Computational pathology models have been shown to depend on the distribution shifts between different domains.

In my thesis, I merely touched upon the topic of domain generalisation, never exploring the methods for aligning the distributions of different datasets or evaluating the effects of putting different datasets into the training or the validation sets.

Zamanitajeddin et al. [201] benchmarked the effectiveness of 30 domain generalisation algorithms on three computational pathology tasks, concluding that self-supervised learning and stain augmentation were consistently among the best-performing algorithms, emphasising their effectiveness in addressing domain shifts.

The dataset I compiled in Chapter 5 brings together data from TCGA, TCIA-CPTAC, DHMC, OUH and four DART sites, meaning that the slides have been stained in at least eight different hospitals and digitised using at least four different scanners. Due to the project requirements, the first batch of images was scanned twice using different scanners. Although the slides were unchanged, the colour palette changed, and the scans of the same slides looked different. One of the possible directions in which the research presented in my thesis could go further is exploring domain generalisation methods, quantifying batch effects, and performing a comprehensive evaluation while placing different parts of the combined dataset into training or test sets.

Additionally, the OUH and DART datasets contained both biopsies and resections. Data coming from a biopsy suffers from more uncertainty in labelling, where more samples can be annotated as malignant, but a subtype cannot be identified due to lack of information on the small tissue size. Comparing the performance of classification algorithms on the biopsies and resection groups might reveal the difference in performance between the two domains. However, collecting a paired (biopsy followed by a resection) dataset and successfully predicting the subtype revealed on the resection directly from a biopsy can potentially result in a change of treatment pathway.

Finally, I collected the desirability of seeing an EVG patch version during the region annotation stage. Although I was not able to present pathologists with an EVG version, it would be interesting to explore how a classifier can make use of both H&E and EVG domains for making predictions.

6.2.2 Multi-stage Classification Pipelines

Pathologists use the presence of adenocarcinoma patterns to determine the presence of adenocarcinomas. But the abundance of slides annotated as adenocarcinomas compared to the limited number of slides with adenocarcinoma pattern annotations can mean that it is easier to learn an adenocarcinoma classifier first, and then learn a separate multi-label classifier to predict the presence of adenocarcinoma patterns when restricted to adenocarcinoma slides. This research direction is supported by results presented in Section 5.5.9. Simpler classifiers that did not explicitly model class dependencies could predict the presence of 5 adenocarcinoma patterns on adenocarcinoma-only slides while being unable to predict all 8 labels simultaneously. Hence, evaluating multi-stage lung cancer classification pipelines and investigating the accumulation of errors presents an interesting path for further research.

6.2.3 Choosing Foundation Models

Chapter 4 showed that it is possible to choose a foundation model for the downstream patch-level classification task using a pretext task at a much lower computational cost. While my attempt at correlating performance on the pretext task of clustering augmented patches was unsuccessful as a method, other researchers explored similar ideas around the same time. Concurrent work has been released by [69] benchmarking whether the latent representations of foundation models were invariant to rotations. Other ideas, like measuring the compactness of patch embeddings as done in metric learning [134], can be explored. I believe that if successful, this method will enable researchers to choose better foundation models for their own tasks at a fraction of computational resources.

6.2.4 Connection with CT

Finally, as the DART dataset keeps growing, I hope that the issues with data anonymisation will be resolved and both pathology and radiology reports will become available to researchers. Then, the original aim of connecting the histology and CT modalities can finally be explored. It has the potential to deepen our understanding of features present in CT scans, improve the quality of diagnosis directly from CT scans, speed up the diagnostic process, and reduce the number of invasive procedures.

6.3 Summary

The research presented in this thesis was a part of the larger DART lung screening project. I believe that my contributions to the annotation of the DART pathology dataset will help researchers continue their work on lung cancer research within and outside the DART project. Furthermore, I hope that other researchers will use the compiled multi-label dataset and my contribution will inspire someone to process and add the TCIA National Lung Screening Trial (TCIA-NLST) data [171] into the mix.

My contributions to optimising the annotation of digitised pathology images by prioritising the annotation of slides and regions from under-represented patterns should help researchers on other projects use their annotation resources more efficiently now that they are equipped with the Ranking Curve AUC metric I proposed for evaluating the ranking methods. I believe that the improved results I achieved by modelling class dependencies and using mixed supervision on partial region annotations and slide-level labels will encourage researchers to collaborate more with domain experts (pathologists, in my case) and incorporate prior knowledge directly into the modelling pipelines.

Finally, I hope that the open-source contributions I made by releasing code for all my published works and the compiled labels for the public datasets, my contributions to TIA-Toolbox [145] library, and my help in releasing the MS-CLAM [179] repository will

encourage the medical imaging community to release data and code with permissive licences more often than it happens now.

Bibliography

[1] NSCLC Targeted Therapy | Non-small Cell Lung Cancer Medication. https://www.cancer.org/cancer/types/lung-cancer/treating-non-small-cell/targeted-therapies.html.

[2] Non-small Cell Lung Cancer Treatment by Stage. https://www.cancer.org/cancer/types/lung-cancer/treating-non-small-cell/by-stage.html.

[3] Pemetrexed (Alimta). https://www.cancerresearchuk.org/about-cancer/treatment/drugs/pemetrexed.

[4] TCGA Dataset. https://www.cancer.gov/tcga.

[5] Treatment for lung cancer. https://www.cancerresearchuk.org/about-cancer/lung-cancer/treatment.

[6] Lung cancer incidence statistics. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer/incidence, May 2015.

[7] Lung cancer survival statistics. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer/survival, May 2015.

[8] CPTAC-LSCC. The Clinical Proteomic Tumor Analysis Consortium Lung Squamous Cell Carcinoma Collection (Version 14) [Data set], 2018.

[9] CPTAC-LUAD. The Clinical Proteomic Tumor Analysis Consortium Lung Adenocarcinoma Collection (Version 12) [Data set], 2018.

[10] CPTAC-LUAD, CPTAC-LSCC. The Clinical Proteomic Tumor Analysis Consortium Lung Adenocarcinoma Collection (Version12) [dataset]. The Clinical Proteomic Tumor Analysis Consortium Lung Squamous Cell Carcinoma Collection (Version14) [dataset], 2018.

[11] H&E stain. *Wikipedia*, Aug. 2021.

[12] Non-Small Cell Lung Cancer Treatment - NCI. https://www.cancer.gov/types/lung/patient/non-small-cell-lung-treatment-pdq, May 2025.

[13] Lung cancer - Treatment. https://www.nhs.uk/conditions/lung-cancer/treatment/, 23 Oct 2017, 3:05 p.m.

[14] A. Aberdeen, N. K. Alham, C. Verrill, and J. Rittscher. AIDA - Annotation of Image Data by Assignments, 2021.

[15] D. R. Aberle, A. M. Adams, C. D. Berg, W. C. Black, J. D. Clapp, R. M. Fagerstrom, I. F. Gareen, C. Gatsonis, P. M. Marcus, and J. D. Sicks. Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, 365(5):395–409, 2011. doi: 10.1056/nejmoa1102873.

[16] K. Abutalip, N. Saeed, M. Khan, and A. E. Saddik. Improving Stain Invariance of CNNs for Segmentation by Fusing Channel Attention and Domain-Adversarial Training. In *Medical Imaging with Deep Learning*, pages 1176–1198. PMLR, Jan. 2024.

[17] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. Sanity Checks for Saliency Maps, Nov. 2020.

[18] S. Agarwal, M. Eltigani Osman Abaker, and O. Daescu. Survival Prediction Based

on Histopathology Imaging and Clinical Data: A Novel, Whole Slide CNN Approach. In M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, Lecture Notes in Computer Science, pages 762–771, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87240-3. doi: 10.1007/978-3-030-87240-3_73.

[19] S. Alfasly, A. Shafique, P. Nejat, J. Khan, A. Alsaafin, G. Alabtah, and H. R. Tizhoosh. Rotation-Agnostic Image Representation Learning for Digital Pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11683–11693, 2024.

[20] N. Alsubaie, M. Shaban, D. R. J. Snead, S. A. Khurram, and N. M. Rajpoot. A Multi-resolution Deep Learning Framework for Lung Adenocarcinoma Growth Pattern Classification. In M. S. Nixon, S. Mahmoodi, and R. Zwiggelaar, editors, *Medical Image Understanding and Analysis - 22nd Conference, MIUA 2018, Southampton, UK, July 9-11, 2018, Proceedings*, volume 894, page 311. Springer, 2018. doi: 10.1007/978-3-319-95921-4_1.

[21] G. Aresta, T. Araújo, S. Kwok, S. S. Chennamsetty, M. Safwan, V. Alex, B. Marami, M. Prastawa, M. Chan, M. Donovan, G. Fernandez, J. Zeineh, M. Kohl, C. Walz, F. Ludwig, S. Braunewell, M. Baust, Q. D. Vu, M. N. N. To, E. Kim, J. T. Kwak, S. Galal, V. Sanchez-Freire, N. Brancati, M. Frucci, D. Riccio, Y. Wang, L. Sun, K. Ma, J. Fang, I. Kone, L. Boulmane, A. Campilho, C. Eloy, A. Polónia, and P. Aguiar. BACH: Grand challenge on breast cancer histology images. *Medical Image Analysis*, 56:122–139, Aug. 2019. ISSN 1361-8415. doi: 10.1016/j.media.2019.05.010.

[22] A. L. Association. PD-L1, PD1,TMB and Lung Cancer. https://www.lung.org/lung-health-diseases/lung-disease-lookup/lung-cancer/symptoms-diagnosis/biomarker-testing/pdl1-pd1-tmb.

[23] M. Aubreville, J. Ganz, J. Ammeling, C. Kaltenecker, and C. Bertram. Model-based Cleaning of the QUILT-1M Pathology Dataset for Text-Conditional Image Synthesis. In *Medical Imaging with Deep Learning*, Apr. 2024.

[24] S. Azizi, L. Culp, J. Freyberg, B. Mustafa, S. Baur, S. Kornblith, T. Chen, N. Tomasev, J. Mitrović, P. Strachan, S. S. Mahdavi, E. Wulczyn, B. Babenko, M. Walker, A. Loh, P.-H. C. Chen, Y. Liu, P. Bavishi, S. M. McKinney, J. Winkens, A. G. Roy, Z. Beaver, F. Ryan, J. Krogue, M. Etemadi, U. Telang, Y. Liu, L. Peng, G. S. Corrado, D. R. Webster, D. Fleet, G. Hinton, N. Houlsby, A. Karthikesalingam, M. Norouzi, and V. Natarajan. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering*, 7(6):756–779, June 2023. ISSN 2157-846X. doi: 10.1038/s41551-023-01049-7.

[25] D. Bahdanau, K. Cho, and Y. Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[26] H. Balata, M. Ruparel, E. O'Dowd, M. Ledson, J. K. Field, S. W. Duffy, S. L. Quaife, A. Sharman, S. Janes, D. Baldwin, R. Booton, and P. A. J. Crosbie. Analysis of the baseline performance of five UK lung cancer screening programmes. *Lung Cancer*, 161:136–140, Nov. 2021. ISSN 0169-5002. doi: 10.1016/j.lungcan.2021.09.012.

[27] D. R. Baldwin, J. Gustafson, L. Pickup, C. Arteta, P. Novotny, J. Declerck, T. Kadir, C. Figueiras, A. Sterba, A. Exell, V. Potesil, P. Holland, H. Spence, A. Clubley, E. O'Dowd, M. Clark, V. Ashford-Turner, M. E. Callister, and F. V. Gleeson. External validation of a convolutional neural network artificial intelligence tool to predict malignancy in pulmonary nodules. *Thorax*, 75(4):306–312, Apr. 2020. ISSN 0040-6376, 1468-3296. doi: 10.1136/thoraxjnl-2019-214104.

[28] D. R. Baldwin, J. Gustafson, L. Pickup, C. Arteta, P. Novotny, J. Declerck, T. Kadir, C. Figueiras, A. Sterba, A. Exell, V. Potesil, P. Holland, H. Spence, A. Clubley, E. ODowd, M. Clark, V. Ashford-Turner, M. E. J. Callister, and F. V. Gleeson. External validation of a convolutional neural network artificial intelligence tool to predict malignancy in pulmonary nodules. *Thorax*, 75(4):306–312, 2020. doi: 10.1136/thoraxjnl-2019-214104.

[29] H. Bao, L. Dong, S. Piao, and F. Wei. BEiT: BERT Pre-Training of Image Transformers. In *International Conference on Learning Representations*, Oct. 2021.

[30] J. A. Barletta, B. Y. Yeap, and L. R. Chirieac. The Prognostic Significance of Grading in Lung Adenocarcinoma. *Cancer*, 116(3):659–669, Feb. 2010. ISSN 0008-543X. doi: 10.1002/cncr.24831.

[31] G. Batchkala, T. Chakraborti, M. McCole, F. Gleeson, and J. Rittscher. Active Data Enrichment by Learning What to Annotate in Digital Pathology. In S. Ali, F. van der Sommen, B. W. Papież, M. van Eijnatten, Y. Jin, and I. Kolenbrander, editors, *Cancer Prevention Through Early Detection*, Lecture Notes in Computer Science, pages 118–127, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-17979-2. doi: 10.1007/978-3-031-17979-2_12.

[32] G. Batchkala, M. Fan, B. Li, M. McCole, C. Brambilla, F. Gleeson, and J. Rittscher. 38 Modelling Class Dependencies for Lung Cancer Subtyping from Digitised Pathology Images. *Lung Cancer*, 190, Apr. 2024. ISSN 0169-5002, 1872-8332. doi: 10.1016/j.lungcan.2024.107599.

[33] G. Batchkala, B. Li, M. Fan, M. McCole, C. Brambilla, F. Gleeson, and J. Rittscher. Accurate Subtyping of Lung Cancers by Modelling Class Dependencies. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, May 2024. doi: 10.1109/ISBI56570.2024.10635232.

[34] G. Batchkala, B. Li, and J. Rittscher. Evaluating Histopathology Foundation

Models for Few-Shot Tissue Clustering: An Application to LC25000 Augmented Dataset Cleaning. In B. Bhattarai, S. Ali, A. Rau, R. Caramalau, A. Nguyen, P. Gyawali, A. Namburete, and D. Stoyanov, editors, *Data Engineering in Medical Imaging*, pages 11–21, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73748-0. doi: 10.1007/978-3-031-73748-0_2.

[35] A. A. Borkowski, M. M. Bui, L. B. Thomas, C. P. Wilson, L. A. DeLand, and S. M. Mastorides. Lung and Colon Cancer Histopathological Image Dataset (LC25000). *arXiv:1912.12142 [cs, eess, q-bio]*, Dec. 2019.

[36] J. Breen, K. Allen, K. Zucker, L. Godson, N. M. Orsi, and N. Ravikumar. A Comprehensive Evaluation of Histopathology Foundation Models for Ovarian Cancer Subtype Classification, Sept. 2024.

[37] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, Aug. 1996. ISSN 1573-0565. doi: 10.1007/BF00058655.

[38] W. Bulten, K. Kartasalo, P.-H. C. Chen, P. Ström, H. Pinckaers, K. Nagpal, Y. Cai, D. F. Steiner, H. van Boven, R. Vink, C. Hulsbergen-van de Kaa, J. van der Laak, M. B. Amin, A. J. Evans, T. van der Kwast, R. Allan, P. A. Humphrey, H. Grönberg, H. Samaratunga, B. Delahunt, T. Tsuzuki, T. Häkkinen, L. Egevad, M. Demkin, S. Dane, F. Tan, M. Valkonen, G. S. Corrado, L. Peng, C. H. Mermel, P. Ruusuvuori, G. Litjens, and M. Eklund. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: The PANDA challenge. *Nature Medicine*, 28(1):154–163, Jan. 2022. ISSN 1546-170X. doi: 10.1038/s41591-021-01620-2.

[39] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, Aug. 2019. ISSN 1546-170X. doi: 10.1038/s41591-019-0508-1.

[40] G. Campanella, R. Kwan, E. Fluder, J. Zeng, A. Stock, B. Veremis, A. D. Polydorides, C. Hedvat, A. Schoenfeld, C. Vanderbilt, P. Kovatch, C. Cordon-Cardo, and T. J. Fuchs. Computational Pathology at Health System Scale – Self-Supervised Foundation Models from Three Billion Images, Oct. 2023.

[41] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[42] A. Chatrian, R. T. Colling, L. Browning, N. K. Alham, K. Sirinukunwattana, S. Malacrino, M. Haghighat, A. Aberdeen, A. Monks, B. Moxley-Wyles, E. Rakha, D. R. J. Snead, J. Rittscher, and C. Verrill. Artificial intelligence for advance requesting of immunohistochemistry in diagnostically uncertain prostate biopsies. *Modern Pathology*, 34(9):1780–1794, Sept. 2021. ISSN 1530-0285. doi: 10.1038/s41379-021-00826-6.

[43] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 40(04):834–848, 2018. ISSN 1939-3539. doi: 10.1109/TPAMI.2017.2699184.

[44] L.-C. Chen, G. Papandreou, I. K. K. Murphy, and A. L. Yuille. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[45] R. J. Chen, M. Y. Lu, M. Shaban, C. Chen, T. Y. Chen, D. F. K. Williamson, and F. Mahmood. Whole Slide Images are 2D Point Clouds: Context-Aware Survival Prediction Using Patch-Based Graph Convolutional Networks. In M. de Bruijne,

P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, Lecture Notes in Computer Science, pages 339–349, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87237-3. doi: 10.1007/978-3-030-87237-3_33.

[46] R. J. Chen, T. Ding, M. Y. Lu, D. F. K. Williamson, G. Jaume, A. H. Song, B. Chen, A. Zhang, D. Shao, M. Shaban, M. Williams, L. Oldenburg, L. L. Weishaupt, J. J. Wang, A. Vaidya, L. P. Le, G. Gerber, S. Sahai, W. Williams, and F. Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, pages 1–13, Mar. 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-02857-3.

[47] R. J. Chen, T. Ding, M. Y. Lu, D. F. K. Williamson, G. Jaume, A. H. Song, B. Chen, A. Zhang, D. Shao, M. Shaban, M. Williams, L. Oldenburg, L. L. Weishaupt, J. J. Wang, A. Vaidya, L. P. Le, G. Gerber, S. Sahai, W. Williams, and F. Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, Mar. 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-02857-3.

[48] S. Chen, G. Campanella, A. Elmas, A. Stock, J. Zeng, A. D. Polydorides, A. J. Schoenfeld, K.-l. Huang, J. Houldsworth, C. Vanderbilt, and T. J. Fuchs. Benchmarking Embedding Aggregation Methods in Computational Pathology: A Clinical Data Perspective. In *Proceedings of the MICCAI Workshop on Computational Pathology*, pages 38–50. PMLR, Nov. 2024.

[49] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607. PMLR, Nov. 2020.

[50] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton. Big Self-

Supervised Models are Strong Semi-Supervised Learners. *arXiv:2006.10029 [cs, stat]*, Oct. 2020.

[51] X. Chen, H. Fan, R. Girshick, and K. He. Improved Baselines with Momentum Contrastive Learning. *arXiv:2003.04297 [cs]*, Mar. 2020.

[52] F. Chollet. Xception: Deep Learning with Depthwise Separable Convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, July 2017. doi: 10.1109/CVPR.2017.195.

[53] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *Journal of Digital Imaging*, 26(6):1045–1057, Dec. 2013. ISSN 1618-727X. doi: 10.1007/s10278-013-9622-7.

[54] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.

[55] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. L. Moreira, N. Razavian, and A. Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine*, 24(10):1559–1567, Oct. 2018. ISSN 1546-170X. doi: 10.1038/s41591-018-0177-5.

[56] CRUK. Biopsy through the skin | Lung cancer | Cancer Research UK. https://www.cancerresearchuk.org/about-cancer/lung-cancer/getting-diagnosed/tests/biopsy-through-skin, .

[57] CRUK. Bronchoscopy for lung cancer | Cancer Research UK.

https://www.cancerresearchuk.org/about-cancer/lung-cancer/getting-diagnosed/tests/bronchoscopy-local-anaesthetic, .

[58] CRUK. Surgical biopsy | Lung cancer | Cancer Research UK. https://www.cancerresearchuk.org/about-cancer/lung-cancer/getting-diagnosed/tests/surgical-biopsy, .

[59] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski. Vision Transformers Need Registers. In *The Twelfth International Conference on Learning Representations*, Oct. 2023.

[60] M. R. Davidson, A. F. Gazdar, and B. E. Clarke. The pivotal role of pathology in the management of lung cancer. *Journal of Thoracic Disease*, 5 Suppl 5:S463–478, Oct. 2013. ISSN 2072-1439. doi: 10.3978/j.issn.2072-1439.2013.08.43.

[61] H. J. de Koning, C. M. van der Aalst, P. A. de Jong, E. T. Scholten, K. Nackaerts, M. A. Heuvelmans, J.-W. J. Lammers, C. Weenink, U. Yousaf-Khan, N. Horeweg, S. van 't Westeinde, M. Prokop, W. P. Mali, F. A. A. M. Hoesein, P. M. A. van Ooijen, J. G. J. V. Aerts, M. A. den Bakker, E. Thunnissen, J. Verschakelen, R. Vliegenthart, J. E. Walter, K. ten Haaf, H. J. M. Groen, and M. Oudkerk. Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. *New England Journal of Medicine*, 382(6):503–513, Feb. 2020. ISSN 0028-4793. doi: 10.1056/NEJMoa1911793.

[62] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[63] J. L. Derks, A.-M. C. Dingemans, R.-J. van Suylen, M. A. den Bakker, R. A. M. Damhuis, E. C. van den Broek, E.-J. Speel, and E. Thunnissen. Is the sum of positive neuroendocrine immunohistochemical stains useful for diagnosis of large cell neuroendocrine carcinoma (LCNEC) on biopsy specimens? *Histopathology*,

74(4):555–566, Mar. 2019. ISSN 0309-0167. doi: 10.1111/his.13800.

[64] K. Desai and J. Johnson. VirTex: Learning Visual Representations From Textual Annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11162–11173, 2021.

[65] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019.

[66] T. DeVries and G. W. Taylor. Improved Regularization of Convolutional Neural Networks with Cutout, Nov. 2017.

[67] T. Ding, S. J. Wagner, A. H. Song, R. J. Chen, M. Y. Lu, A. Zhang, A. J. Vaidya, G. Jaume, M. Shaban, A. Kim, D. F. K. Williamson, B. Chen, C. Almagro-Perez, P. Doucet, S. Sahai, C. Chen, D. Komura, A. Kawabe, S. Ishikawa, G. Gerber, T. Peng, L. P. Le, and F. Mahmood. Multimodal Whole Slide Foundation Model for Pathology, Nov. 2024.

[68] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021.

[69] M. Elphick, S. Turajlic, and G. Yang. Are the Latent Representations of Foundation Models for Pathology Invariant to Rotation?, Dec. 2024.

[70] M. Everingham, L. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.*, 88(2): 303–338, 2010.

[71] M. Fan, T. Chakraborti, E. I. Chang, Y. Xu, and J. Rittscher. Microscopic Fine-Grained Instance Classification Through Deep Attention. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lec-*

ture Notes in Bioinformatics), 12265 LNCS:490–499, 2020. ISSN 16113349. doi: 10.1007/978-3-030-59722-1_47.

[72] F. Farjah, S. E. Monsell, R. T. Greenlee, M. K. Gould, R. Smith-Bindman, M. P. Banegas, K. Schoen, A. Ramaprasan, and D. S. Buist. Patient and Nodule Characteristics Associated With a Lung Cancer Diagnosis Among Individuals With Incidentally Detected Lung Nodules. *Chest*, 163(3):719–730, Mar. 2023. ISSN 0012-3692. doi: 10.1016/j.chest.2022.09.030.

[73] A. Filiot, R. Ghermi, A. Olivier, P. Jacob, L. Fidon, A. Mac Kain, C. Saillard, and J.-B. Schiratti. Scaling Self-Supervised Learning for Histopathology with Masked Image Modeling, July 2023.

[74] A. Filiot, P. Jacob, A. M. Kain, and C. Saillard. Phikon-v2, A large and public feature extractor for biomarker prediction, Sept. 2024.

[75] E. B. Fowlkes and C. L. Mallows. A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*, 78(383):553–569, Sept. 1983. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.1983.10478008.

[76] Y. Gal and Z. Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In M.-F. Balcan and K. Q. Weinberger, editors, *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48, pages 1050–1059. JMLR.org, 2016.

[77] A. Gallagher-Syed, E. Pontarini, M. J. Lewis, M. R. Barnes, and G. Slabaugh. Going Beyond H&E and Oncology: How Do Histopathology Foundation Models Perform for Multi-stain IHC and Immunology?, Oct. 2024.

[78] J. Gamper and N. Rajpoot. ARCH Dataset, 2021.

[79] J. Gamper and N. Rajpoot. Multiple Instance Captioning: Learning Repre-

sentations From Histopathology Textbooks and Articles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16549–16559, June 2021.

[80] A. Gholami, K. Kwon, B. Wu, Z. Tai, X. Yue, P. Jin, S. Zhao, and K. Keutzer. SqueezeNext: Hardware-Aware Neural Network Design, Aug. 2018.

[81] F. Gleeson and A. Powell. DART: The Integration and Analysis of Data using Artificial Intelligence to Improve Patient Outcomes with Thoracic Diseases, 2020.

[82] L. C. Group. Small Cell Lung Cancer | Diagnosing and Treating SCLC, June 2025.

[83] S. Gulati, T. Ivic-Pavlicic, J. Joasil, R. Flores, and E. Taioli. Outcomes in Incidentally Versus Screening Detected Stage I Lung Cancer Surgery Patients. *Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer*, 19(4):581–588, Apr. 2024. ISSN 1556-1380. doi: 10.1016/j.jtho.2023.11.008.

[84] H. R. Gwon, A. La Woo, S. H. Yong, Y. Park, S. Y. Kim, E. Y. Kim, J. Y. Jung, Y. A. Kang, M. S. Park, S. Y. Park, and S. H. Lee. Factors affecting accuracy of clinical staging in resectable non-small cell lung cancer in a real-world study. *Thoracic Cancer*, 15(9):730–737, Feb. 2024. ISSN 1759-7706. doi: 10.1111/1759-7714.15253.

[85] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality Reduction by Learning an Invariant Mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742, June 2006. doi: 10.1109/CVPR.2006.100.

[86] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[87] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum Contrast for Unsuper-

vised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[88] M. B. Heldwein, G. Schlachtenberger, F. Doerr, H. Menghesha, G. Bennink, K.-M. Schroeder, S. C. Schaefer, T. Wahlers, and K. Hekmat. Different pulmonary adenocarcinoma growth patterns significantly affect survival. *Surgical Oncology*, 40:101674, Mar. 2022. ISSN 1879-3320. doi: 10.1016/j.suronc.2021.101674.

[89] G. Hinton, O. Vinyals, and J. Dean. Distilling the Knowledge in a Neural Network, Mar. 2015.

[90] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors, July 2012.

[91] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam. Searching for MobileNetV3, Nov. 2019.

[92] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, Apr. 2017.

[93] J. Hu, L. Shen, and G. Sun. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[94] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

[95] T. Huang, J. Li, C. Zhang, Q. Hong, D. Jiang, M. Ye, and S. Duan. Distinguishing Lung Adenocarcinoma from Lung Squamous Cell Carcinoma by Two Hypomethy-

lated and Three Hypermethylated Genes: A Meta-Analysis. *PLOS ONE*, 11(2): e0149088, Feb. 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0149088.

[96] Z. Huang, F. Bianchi, M. Yuksekgonul, T. Montine, and J. Zou. Leveraging medical Twitter to build a visual–language foundation model for pathology AI, Apr. 2023.

[97] Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine, and J. Zou. A visual–language foundation model for pathology image analysis using medical Twitter. *Nature Medicine*, 29(9):2307–2316, Sept. 2023. ISSN 1546-170X. doi: 10.1038/s41591-023-02504-3.

[98] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1): 193–218, Dec. 1985. ISSN 0176-4268, 1432-1343. doi: 10.1007/BF01908075.

[99] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size, Nov. 2016.

[100] W. IHC. EVG stain. http://www.ihcworld.com/_protocols/special_stains/vvg.htm.

[101] W. O. Ikezogwo, M. S. Seyfioglu, F. Ghezloo, D. S. C. Geva, F. S. Mohammed, P. K. Anand, R. Krishna, and L. Shapiro. Quilt-1M: One Million Image-Text Pairs for Histopathology. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, Nov. 2023.

[102] M. Ilse, J. Tomczak, and M. Welling. Attention-based Deep Multiple Instance Learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2127–2136. PMLR, July 2018.

[103] S. Ioffe and C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 448–456, Lille, France, Jan. 2015. PMLR.

[104] G. Jaume, L. Oldenburg, A. Vaidya, R. J. Chen, D. F. K. Williamson, T. Peeters, A. H. Song, and F. Mahmood. Transcriptomics-guided Slide Representation Learning in Computational Pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9632–9644, 2024.

[105] M. Joerger, A. Omlin, T. Cerny, and M. Früh. The role of pemetrexed in advanced non small-cell lung cancer: Special focus on pharmacology and mechanism of action. *Current Drug Targets*, 11(1):37–47, Jan. 2010. ISSN 1873-5592. doi: 10.2174/138945010790030974.

[106] N. Karachaliou, M. Fernandez-Bruno, and R. Rosell. Strategies for first-line immunotherapy in squamous cell lung cancer: Are combinations a game changer? *Translational Lung Cancer Research*, 7(Suppl 3):S198–S201, Sept. 2018. ISSN 2218-6751. doi: 10.21037/tlcr.2018.07.02.

[107] A. Karpathy. What I learned from competing against a ConvNet on ImageNet. http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/, 2014.

[108] J. N. Kather, N. Halama, and A. Marx. 100,000 histological images of human colorectal cancer and healthy tissue, Apr. 2018.

[109] Y. J. Kim, H. Jang, K. Lee, S. Park, S.-G. Min, C. Hong, J. H. Park, K. Lee, J. Kim, W. Hong, H. Jung, Y. Liu, H. Rajkumar, M. Khened, G. Krishnamurthi, S. Yang, X. Wang, C. H. Han, J. T. Kwak, J. Ma, Z. Tang, B. Marami, J. Zeineh, Z. Zhao, P.-A. Heng, R. Schmitz, F. Madesta, T. Rösch, R. Werner, J. Tian, E. Puybareau, M. Bovio, X. Zhang, Y. Zhu, S. Y. Chun, W.-K. Jeong, P. Park, and J. Choi. PAIP 2019: Liver cancer segmentation challenge. *Medical Image Analysis*, 67:101854, Jan. 2021. ISSN 1361-8415. doi: 10.1016/j.media.2020.101854.

[110] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Represen-*

tations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[111] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a Meeting Held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1106–1114, 2012.

[112] E. Kuhn, P. Morbini, A. Cancellieri, S. Damiani, A. Cavazza, and CE. Comin. Adenocarcinoma classification: Patterns and prognosis. *Pathologica-Journal of the Italian Society of Anatomic Pathology and Diagnostic Cytopathology*, 110(1): 5–11, 2018.

[113] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, Mar. 1955. ISSN 0028-1441, 1931-9193. doi: 10.1002/nav.3800020109.

[114] K. Lami, A. Bychkov, K. Matsumoto, R. Attanoos, S. Berezowska, L. Brcic, A. Cavazza, J. C. English, A. T. Fabro, K. Ishida, Y. Kashima, B. T. Larsen, A. M. Marchevsky, T. Miyazaki, S. Morimoto, A. C. Roden, F. Schneider, M. Soshi, M. L. Smith, K. Tabata, A. M. Takano, K. Tanaka, T. Tanaka, T. Tsuchiya, T. Nagayasu, and J. Fukuoka. Overcoming the Interobserver Variability in Lung Adenocarcinoma Subtyping: A Clustering Approach to Establish a Ground Truth for Downstream Applications. *Archives of Pathology & Laboratory Medicine*, 147(8):885–895, Nov. 2022. ISSN 0003-9985. doi: 10.5858/arpa.2022-0051-OA.

[115] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov. 1998. ISSN 1558-2256. doi: 10.1109/5.726791.

[116] J. E. Leeman, A. Rimner, J. Montecalvo, M. Hsu, Z. Zhang, D. von Reibnitz, K. Panchoo, E. Yorke, P. S. Adusumilli, W. Travis, and A. J. Wu. Histologic subtype in core lung biopsies of early-stage lung adenocarcinoma is a prognostic factor for treatment response and failure patterns after stereotactic body radiation therapy. *International journal of radiation oncology, biology, physics*, 97(1):138–145, Jan. 2017. ISSN 0360-3016. doi: 10.1016/j.ijrobp.2016.09.037.

[117] B. Li, Y. Li, and K. W. Eliceiri. Dual-Stream Multiple Instance Learning Network for Whole Slide Image Classification With Self-Supervised Contrastive Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2021.

[118] H. Li, F. Yang, Y. Zhao, X. Xing, J. Zhang, M. Gao, J. Huang, L. Wang, and J. Yao. DT-MIL: Deformable Transformer for Multi-instance Learning on Histopathological Image. In M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, Lecture Notes in Computer Science, pages 206–216, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87237-3. doi: 10.1007/978-3-030-87237-3_20.

[119] M. Lin, Q. Chen, and S. Yan. Network In Network, Mar. 2014.

[120] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1. doi: 10.1007/978-3-319-10602-1_48.

[121] J. Lipkova, T. Y. Chen, M. Y. Lu, R. J. Chen, M. Shady, M. Williams, J. Wang, Z. Noor, R. N. Mitchell, M. Turan, G. Coskun, F. Yilmaz, D. Demir, D. Nart, K. Basak, N. Turhan, S. Ozkara, Y. Banz, K. E. Odening, and F. Mahmood. Deep learning-enabled assessment of cardiac allograft rejection from endomyocardial

biopsies. *Nature Medicine*, 28(3):575–582, Mar. 2022. ISSN 1546-170X. doi: 10.1038/s41591-022-01709-2.

[122] G. Litjens, P. Bandi, B. Ehteshami Bejnordi, O. Geessink, M. Balkenhol, P. Bult, A. Halilovic, M. Hermsen, R. van de Loo, R. Vogels, Q. F. Manson, N. Stathonikos, A. Baidoshvili, P. van Diest, C. Wauters, M. van Dijk, and J. van der Laak. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: The CAME-LYON dataset. *GigaScience*, 7(6):giy065, June 2018. ISSN 2047-217X. doi: 10.1093/gigascience/giy065.

[123] I. Loshchilov and F. Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts. In *International Conference on Learning Representations*, 2017.

[124] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering 2021 5:6*, pages 555–570, Mar. 2021. ISSN 2157-846X. doi: 10.1038/s41551-020-00682-w.

[125] M. Y. Lu, B. Chen, D. F. K. Williamson, R. J. Chen, I. Liang, T. Ding, G. Jaume, I. Odintsov, L. P. Le, G. Gerber, A. V. Parwani, A. Zhang, and F. Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, Mar. 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-02856-4.

[126] W. Lu, M. Toss, M. Dawood, E. Rakha, N. Rajpoot, and F. Minhas. SlideGraph+: Whole slide image level graphs to predict HER2 status in breast cancer. *Medical Image Analysis*, 80:102486, Aug. 2022. ISSN 1361-8415. doi: 10.1016/j.media.2022.102486.

[127] S. Mangal, A. Chaurasia, and A. Khajanchi. Convolution Neural Networks for diagnosing colon and lung cancer histopathological images. *ArXiv*, Sept. 2020.

[128] N. Marini, S. Otálora, F. Ciompi, G. Silvello, S. Marchesin, S. Vatrano, G. Buttafuoco, M. Atzori, and H. Müller. Multi-Scale Task Multiple Instance Learning for

the Classification of Digital Pathology Images with Global Annotations. In *Proceedings of the MICCAI Workshop on Computational Pathology*, pages 170–181. PMLR, Sept. 2021.

[129] M. Masud, N. Sikder, A.-A. Nahid, A. K. Bairagi, and M. A. AlZain. A Machine Learning Approach to Diagnosing Lung and Colon Cancer Using a Deep Learning-Based Classification Framework. *Sensors*, 21(3):748, Jan. 2021. ISSN 1424-8220. doi: 10.3390/s21030748.

[130] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, Sept. 2020.

[131] A. McWilliams, M. C. Tammemagi, J. R. Mayo, H. Roberts, G. Liu, K. Soghrati, K. Yasufuku, S. Martel, F. Laberge, M. Gingras, S. Atkar-Khattra, C. D. Berg, K. Evans, R. Finley, J. Yee, J. English, P. Nasute, J. Goffin, S. Puksa, L. Stewart, S. Tsai, M. R. Johnston, D. Manos, G. Nicholas, G. D. Goss, J. M. Seely, K. Amjadi, A. Tremblay, P. Burrowes, P. MacEachern, R. Bhatia, M.-S. Tsao, and S. Lam. Probability of Cancer in Pulmonary Nodules Detected on First Screening CT. *The New England journal of medicine*, 369(10):910–919, Sept. 2013. ISSN 0028-4793. doi: 10.1056/NEJMoa1214726.

[132] S. Mehmood, T. M. Ghazal, M. A. Khan, M. Zubair, M. T. Naseem, T. Faiz, and M. Ahmad. Malignancy Detection in Lung and Colon Histopathology Images Using Transfer Learning With Class Selective Image Processing. *IEEE Access*, 10: 25657–25668, 2022. ISSN 2169-3536. doi: 10.1109/ACCESS.2022.3150924.

[133] R. Meza, C. Meernik, J. Jeon, and M. L. Cote. Lung Cancer Incidence Trends by Gender, Race and Histology in the United States, 1973–2010. *PLOS ONE*, 10(3): 1–14, Jan. 2015. doi: 10.1371/journal.pone.0121323.

[134] K. Musgrave, S. Belongie, and S.-N. Lim. A Metric Learning Reality Check. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK,*

August 23–28, 2020, Proceedings, Part XXV, pages 681–699, Berlin, Heidelberg, Aug. 2020. Springer-Verlag. ISBN 978-3-030-58594-5. doi: 10.1007/978-3-030-58595-2_41.

[135] A. Myronenko, Z. Xu, D. Yang, H. R. Roth, and D. Xu. Accounting for Dependencies in Deep Learning Based Multiple Instance Learning for Whole Slide Imaging. In M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, Lecture Notes in Computer Science, pages 329–338, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87237-3. doi: 10.1007/978-3-030-87237-3_32.

[136] D. Nechaev, A. Pchelnikov, and E. Ivanova. Hibou: A Family of Foundational Vision Transformers for Pathology, June 2024.

[137] P. Neidlinger, O. S. M. E. Nahhas, H. S. Muti, T. Lenz, M. Hoffmeister, H. Brenner, M. van Treeck, R. Langer, B. Dislich, H. M. Behrens, C. Röcken, S. Foersch, D. Truhn, A. Marra, O. L. Saldanha, and J. N. Kather. Benchmarking foundation models as feature extractors for weakly-supervised computational pathology, Dec. 2024.

[138] P. Ngo, W. A. Cooper, S. Wade, K. M. Fong, K. Canfell, D. Karikios, and M. Weber. Why PD-L1 expression varies between studies of lung cancer: Results from a Bayesian meta-analysis. *Scientific Reports*, 15(1):4166, Feb. 2025. ISSN 2045-2322. doi: 10.1038/s41598-024-80301-9.

[139] A. G. Nicholson, K. Torkko, P. Viola, E. Duhig, K. Geisinger, A. C. Borczuk, K. Hiroshima, M. S. Tsao, A. Warth, S. Lantuejoul, P. A. Russell, E. Thunnissen, A. Marchevsky, M. Mino-Kenudson, M. B. Beasley, J. Botling, S. Dacic, Y. Yatabe, M. Noguchi, W. D. Travis, K. Kerr, F. R. Hirsch, L. R. Chirieac, I. I. Wistuba, A. Moreira, J.-H. Chung, T. Y. Chou, L. Bubendorf, G. Chen, G. Pelosi, C. Poleri, F. C. Detterbeck, and W. A. Franklin. Interobserver Variation Among Pathologists

And Refinement Of Criteria In Distinguishing Separate Primary Tumours From Intrapulmonary Metastases In Lung. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer*, 13(2):205–217, Feb. 2018. ISSN 1556-0864. doi: 10.1016/j.jtho.2017.10.019.

[140] A. G. Nicholson, M. S. Tsao, M. B. Beasley, A. C. Borczuk, E. Brambilla, W. A. Cooper, S. Dacic, D. Jain, K. M. Kerr, S. Lantuejoul, M. Noguchi, M. Papotti, N. Rekhtman, G. Scagliotti, P. van Schil, L. Sholl, Y. Yatabe, A. Yoshida, and W. D. Travis. The 2021 WHO Classification of Lung Tumors: Impact of Advances Since 2015. *Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer*, 17(3):362–387, Mar. 2022. ISSN 1556-1380. doi: 10.1016/j.jtho.2021.11.003.

[141] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, Feb. 2024.

[142] D. C. Paech, A. R. Weston, N. Pavlakis, A. Gill, N. Rajan, H. Barraclough, B. Fitzgerald, and M. Van Kooten. A systematic review of the interobserver variability for histology in the differentiation between squamous and nonsquamous non-small cell lung cancer. *Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer*, 6(1):55–63, Jan. 2011. ISSN 1556-1380. doi: 10.1097/JTO.0b013e3181fc0878.

[143] J. Park, J. Lee, Y. J. Jeon, S. Shin, J. H. Cho, H.-K. Kim, Y. S. Choi, J. Kim, J. I. Zo, and Y. M. Shim. Adjuvant Chemotherapy in Patients with Node-Negative Non-Small Cell Lung Cancer with Satellite Pulmonary Nodules in the Same Lobe. *Journal of Chest Surgery*, 55(1):10–19, Feb. 2022. ISSN 2765-1606. doi: 10.5090/jcs.21.110.

[144] N. Parry. How Histology Slides are Prepared, Jan. 2020.

[145] J. Pocock, S. Graham, Q. D. Vu, M. Jahanifar, S. Deshpande, G. Hadjigeorghiou, A. Shephard, R. M. S. Bashir, M. Bilal, W. Lu, D. Epstein, F. Minhas, N. M. Rajpoot, and S. E. A. Raza. TIAToolbox as an end-to-end library for advanced tissue image analytics. *Communications Medicine*, 2(1):1–14, Sept. 2022. ISSN 2730-664X. doi: 10.1038/s43856-022-00186-5.

[146] S. F. Powell and A. Z. Dudek. Tailoring treatment of nonsmall cell lung cancer by tissue type: Role of pemetrexed. *Pharmacogenomics and personalized medicine*, 2:21–37, June 2009. ISSN 1178-7066. doi: 10.2147/pgpm.s3977.

[147] B. Qiao, K. Jumai, J. Ainiwaer, M. Niyaz, Y. Zhang, Y. Ma, L. Zhang, W. Luh, and I. Sheyhidin. A novel transfer-learning based physician-level general and subtype classifier for non-small cell lung cancer. *Heliyon*, 8(12):e11981, Dec. 2022. ISSN 2405-8440. doi: 10.1016/j.heliyon.2022.e11981.

[148] J. Qiu, M. Aubreville, F. Wilm, M. Öttl, J. Utz, M. Schlereth, and K. Breininger. Leveraging Image Captions for Selective Whole Slide Image Annotation. In M. G. Linguraru, Q. Dou, A. Feragen, S. Giannarou, B. Glocker, K. Lekadir, and J. A. Schnabel, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 207–217, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72390-2. doi: 10.1007/978-3-031-72390-2_20.

[149] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, July 2021.

[150] W. M. Rand. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846–850, Dec. 1971. ISSN 0162-

1459, 1537-274X. doi: 10.1080/01621459.1971.10482356.

[151] T. R. Rasmussen, A. Gouliaev, E. Jakobsen, K. Hjorthaug, L. U. Larsen, P. Meldgaard, J. Thygesen, R. Bibi, L. B. Møller, A. Arshad, B. Folkersen, A. Højsgaard, Z. Saghir, K. R. Larsen, and J. Ravn. Impact of multidisciplinary team discrepancies on comparative lung cancer outcome analyses and treatment equality. *BMC Cancer*, 24:1423, Nov. 2024. ISSN 1471-2407. doi: 10.1186/s12885-024-13188-4.

[152] M. Raza, S. Bashir, T. Qaiser, and N. Rajpoot. Stain-Invariant Representation for Tissue Classification in Histology Images, Nov. 2024.

[153] S. Ricciardi, S. Tomao, and F. de Marinis. Pemetrexed as first-line therapy for non-squamous non-small cell lung cancer. *Therapeutics and Clinical Risk Management*, 5:781–787, 2009. ISSN 1176-6336. doi: 10.2147/tcrm.s3195.

[154] A. Rosenberg and J. Hirschberg. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In *Conference on Empirical Methods in Natural Language Processing*, June 2007.

[155] S. Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[156] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, Dec. 2015. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-015-0816-y.

[157] P. A. Russell and G. M. Wright. Predominant histologic subtype in lung adenocarcinoma predicts benefit from adjuvant chemotherapy in completely resected patients: Discovery of a holy grail? *Annals of Translational Medicine*, 4(1):16, Jan. 2016. ISSN 2305-5839. doi: 10.3978/j.issn.2305-5839.2015.10.21.

[158] C. Saillard, R. Jenatton, F. Llinares-López, Z. Mariet, D. Cahané, E. Durand, and J.-P. Vert. H-optimus-0. https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0, 2024.

[159] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks, Mar. 2019.

[160] G. Shaikovski, A. Casson, K. Severson, E. Zimmermann, Y. K. Wang, J. D. Kunz, J. A. Retamero, G. Oakley, D. Klimstra, C. Kanan, M. Hanna, M. Zelechowski, J. Viret, N. Tenenholtz, J. Hall, N. Fusi, R. Yousfi, P. Hamilton, W. A. Moye, E. Vorontsov, S. Liu, and T. J. Fuchs. PRISM: A Multi-Modal Generative Foundation Model for Slide-Level Histopathology, May 2024.

[161] D. Shao, R. J. Chen, A. H. Song, J. Runevic, M. Y. Lu, T. Ding, and F. Mahmood. Do Multiple Instance Learning Models Transfer? In *Forty-Second International Conference on Machine Learning*, June 2025.

[162] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, and y. zhang. TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification. In *Advances in Neural Information Processing Systems*, volume 34, pages 2136–2147. Curran Associates, Inc., 2021.

[163] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*, 2015.

[164] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

[165] A. Stang, H. Pohlabeln, K. M. Müller, I. Jahn, K. Giersiepen, and K.-H. Jöckel. Diagnostic agreement in the histopathological evaluation of lung cancer tissue in a

population-based case-control study. *Lung Cancer (Amsterdam, Netherlands)*, 52 (1):29–36, Apr. 2006. ISSN 0169-5002. doi: 10.1016/j.lungcan.2005.11.012.

[166] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going Deeper With Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[167] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, June 2016. doi: 10.1109/CVPR.2016.308.

[168] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), Feb. 2017. ISSN 2374-3468.

[169] M. A. Talukder, M. M. Islam, M. A. Uddin, A. Akhter, K. F. Hasan, and M. A. Moni. Machine learning-based lung and colon cancer detection using deep feature extraction and ensemble learning. *Expert Systems with Applications*, 205:117695, Nov. 2022. ISSN 09574174. doi: 10.1016/j.eswa.2022.117695.

[170] M. Tan and Q. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 6105–6114. PMLR, Jan. 2019.

[171] N. L. S. T. R. Team. Data from the national lung screening trial (NLST) [Data set]. Available online at The Cancer Imaging Archive, 2013.

[172] D. Tellez, M. Balkenhol, N. Karssemeijer, G. Litjens, J. van der Laak, and F. Ciompi. H and E stain augmentation improves generalization of convolutional

networks for histopathological mitosis detection. In *Medical Imaging 2018: Digital Pathology*, volume 10581, page 105810Z, Mar. 2018. doi: 10.1117/12.2293048.

[173] n. The National Lung Screening Trial Research Team. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. *New England Journal of Medicine*, 365(5):395–409, Aug. 2011. ISSN 0028-4793. doi: 10.1056/NEJMoa1102873.

[174] M. Toğaçar. Disease type detection in lung and colon cancer images using the complement approach of inefficient sets. *Computers in Biology and Medicine*, 137: 104827, Oct. 2021. ISSN 00104825. doi: 10.1016/j.compbiomed.2021.104827.

[175] P. Tomasini, F. Barlesi, C. Mascaux, and L. Greillier. Pemetrexed for advanced stage nonsquamous non-small cell lung cancer: Latest evidence about its extended use and outcomes. *Therapeutic Advances in Medical Oncology*, 8(3):198–208, May 2016. ISSN 1758-8340. doi: 10.1177/1758834016644155.

[176] N. Tomita, B. Abdollahi, J. Wei, B. Ren, A. Suriawinata, and S. Hassanpour. Attention-Based Deep Neural Networks for Detection of Cancerous and Precancerous Esophagus Tissue on Histopathological Slides. *JAMA Network Open*, 2 (11):e1914645, Nov. 2019. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2019.14645.

[177] L. A. Torre, R. L. Siegel, and A. Jemal. Lung Cancer Statistics. In *Lung Cancer and Personalized Medicine: Current Knowledge and Therapies*, pages 1–19. Springer International Publishing, Cham, 2016. ISBN 978-3-319-24223-1. doi: 10.1007/978-3-319-24223-1_1.

[178] P. Tourniaire, M. Ilie, P. Hofman, N. Ayache, and H. Delingette. Attention-based Multiple Instance Learning with Mixed Supervision on the Camelyon16 Dataset. In *Proceedings of the MICCAI Workshop on Computational Pathology*, pages 216–226. PMLR, Sept. 2021.

[179] P. Tourniaire, M. Ilie, P. Hofman, N. Ayache, and H. Delingette. MS-CLAM: Mixed supervision for the classification and localization of tumors in Whole Slide Images. *Medical Image Analysis*, 85:102763, Apr. 2023. ISSN 1361-8415. doi: 10.1016/j.media.2023.102763.

[180] W. D. Travis, E. Brambilla, A. G. Nicholson, Y. Yatabe, J. H. M. Austin, M. B. Beasley, L. R. Chirieac, S. Dacic, E. Duhig, D. B. Flieder, K. Geisinger, F. R. Hirsch, Y. Ishikawa, K. M. Kerr, M. Noguchi, G. Pelosi, C. A. Powell, M. S. Tsao, and I. Wistuba. The 2015 World Health Organization Classification of Lung tumours: Impact of Genetic, Clinical and Radiologic Advances Since the 2004 Classification. *Journal of Thoracic Oncology*, 10(9):1243–1260, 2015. ISSN 1556-0864. doi: 10.1097/JTO.0000000000000630.

[181] C. Trinidad López, C. Delgado Sánchez-Gracián, E. Utrera Pérez, C. Jurado Basildo, and C. A. Sepúlveda Villegas. Incidental pulmonary nodules: Characterization and management. *Radiología (English Edition)*, 61(5):357–369, Sept. 2019. ISSN 2173-5107. doi: 10.1016/j.rxeng.2019.06.002.

[182] J. van der Laak, G. Litjens, and F. Ciompi. Deep learning in histopathology: The path to the clinic. *Nature Medicine*, 27(5):775–784, May 2021. ISSN 1546-170X. doi: 10.1038/s41591-021-01343-4.

[183] L. van Eekelen, J. Spronck, M. Looijen-Salamon, S. Vos, E. Munari, I. Girolami, A. Eccher, B. Acs, C. Boyaci, G. S. de Souza, M. Demirel-Andishmand, L. D. Meesters, D. Zegers, L. van der Woude, W. Theelen, M. van den Heuvel, K. Grünberg, B. van Ginneken, J. van der Laak, and F. Ciompi. Comparing deep learning and pathologist quantification of cell-level PD-L1 expression in non-small cell lung cancer whole-slide images. *Scientific Reports*, 14(1):7136, Mar. 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-57067-1.

[184] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is All you Need. In I. Guyon, U. von Luxburg, S. Ben-

gio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

[185] E. Vorontsov, A. Bozkurt, A. Casson, G. Shaikovski, M. Zelechowski, K. Severson, E. Zimmermann, J. Hall, N. Tenenholtz, N. Fusi, E. Yang, P. Mathieu, A. van Eck, D. Lee, J. Viret, E. Robert, Y. K. Wang, J. D. Kunz, M. C. H. Lee, J. H. Bernhard, R. A. Godrich, G. Oakley, E. Millar, M. Hanna, H. Wen, J. A. Retamero, W. A. Moye, R. Yousfi, C. Kanan, D. S. Klimstra, B. Rothrock, S. Liu, and T. J. Fuchs. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, pages 1–12, July 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-03141-0.

[186] X. Wang, S. Yang, J. Zhang, M. Wang, J. Zhang, J. Huang, W. Yang, and X. Han. TransPath: Transformer-Based Self-supervised Learning for Histopathological Image Classification. In M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, Lecture Notes in Computer Science, pages 186–195, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87237-3. doi: 10.1007/978-3-030-87237-3_18.

[187] X. Wang, S. Yang, J. Zhang, M. Wang, J. Zhang, W. Yang, J. Huang, and X. Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 81:102559, Oct. 2022. ISSN 1361-8415. doi: 10.1016/j.media.2022.102559.

[188] X. Wang, J. Zhao, E. Marostica, W. Yuan, J. Jin, J. Zhang, R. Li, H. Tang, K. Wang, Y. Li, F. Wang, Y. Peng, J. Zhu, J. Zhang, C. R. Jackson, J. Zhang, D. Dillon, N. U. Lin, L. Sholl, T. Denize, D. Meredith, K. L. Ligon, S. Signoretti, S. Ogino, J. A. Golden, M. P. Nasrallah, X. Han, S. Yang, and K.-H. Yu. A pathology foundation

model for cancer diagnosis and prognosis prediction. *Nature*, 634(8035):970–978, Oct. 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07894-z.

[189] Z. Wang, J. Li, Z. Pan, W. Li, A. Sisk, H. Ye, W. Speier, and C. W. Arnold. Hierarchical Graph Pathomic Network for Progression Free Survival Prediction. In M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, Lecture Notes in Computer Science, pages 227–237, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87237-3. doi: 10.1007/978-3-030-87237-3_22.

[190] J. W. Wei, L. J. Tafe, Y. A. Linnik, L. J. Vaickus, N. Tomita, and S. Hassanpour. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. *Scientific Reports*, 9(1):3358, Mar. 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-40041-7.

[191] G. Wölflein, D. Ferber, A. R. Meneghetti, O. S. M. E. Nahhas, D. Truhn, Z. I. Carrero, D. J. Harrison, O. Arandjelović, and J. N. Kather. Benchmarking Pathology Feature Extractors for Whole Slide Image Classification, June 2024.

[192] R. Wood, E. Domingo, K. Sirinukunwattana, M. W. Lafarge, V. H. Koelzer, T. S. Maughan, and J. Rittscher. Joint Prediction of Response to Therapy, Molecular Traits, and Spatial Organisation in Colorectal Cancer Biopsies. In H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, and R. Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 758–767, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43904-9. doi: 10.1007/978-3-031-43904-9_73.

[193] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He. Aggregated Residual Transformations for Deep Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1492–1500, 2017.

[194] H. Xu, N. Usuyama, J. Bagga, S. Zhang, R. Rao, T. Naumann, C. Wong, Z. Gero, J. González, Y. Gu, Y. Xu, M. Wei, W. Wang, S. Ma, F. Wei, J. Yang, C. Li, J. Gao, J. Rosemon, T. Bower, S. Lee, R. Weerasinghe, B. J. Wright, A. Robicsek, B. Piening, C. Bifulco, S. Wang, and H. Poon. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, June 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07441-w.

[195] H. Xu, N. Usuyama, J. Bagga, S. Zhang, R. Rao, T. Naumann, C. Wong, Z. Gero, J. González, Y. Gu, Y. Xu, M. Wei, W. Wang, S. Ma, F. Wei, J. Yang, C. Li, J. Gao, J. Rosemon, T. Bower, S. Lee, R. Weerasinghe, B. J. Wright, A. Robicsek, B. Piening, C. Bifulco, S. Wang, and H. Poon. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, June 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07441-w.

[196] D. Yaldız, A. Acar, Ş. Örs Kaya, Z. Aydoğdu, S. Gürsoy, and S. Yaldız. Papillary predominant histological subtype predicts poor survival in lung adenocarcinoma. *Turkish Journal of Thoracic and Cardiovascular Surgery*, 27(3):360–366, June 2019. ISSN 1301-5680. doi: 10.5606/tgkdc.dergisi.2019.17284.

[197] H. Yang, L. Chen, Z. Cheng, M. Yang, J. Wang, C. Lin, Y. Wang, L. Huang, Y. Chen, S. Peng, Z. Ke, and W. Li. Deep learning-based six-type classifier for lung cancer and mimics from histopathological whole slide images: A retrospective study. *BMC Medicine*, 19(1):80, Mar. 2021. ISSN 1741-7015. doi: 10.1186/s12916-021-01953-2.

[198] F. Yu and V. Koltun. Multi-Scale Context Aggregation by Dilated Convolutions. In Y. Bengio and Y. LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[199] G. Yu, K. Sun, C. Xu, X.-H. Shi, C. Wu, T. Xie, R.-Q. Meng, X.-H. Meng, K.-S. Wang, H.-M. Xiao, and H.-W. Deng. Accurate recognition of colorec-

tal cancer with semi-supervised deep learning on pathological images. *Nature Communications*, 12(1):6311, Nov. 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-26643-8.

[200] H. Yu, T. A. Boyle, C. Zhou, D. L. Rimm, and F. R. Hirsch. PD-L1 Expression in Lung Cancer. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer*, 11(7):964–975, July 2016. ISSN 1556-0864. doi: 10.1016/j.jtho.2016.04.014.

[201] N. Zamanitajeddin, M. Jahanifar, K. Xu, F. Siraj, and N. Rajpoot. Benchmarking Domain Generalization Algorithms in Computational Pathology, Sept. 2024.

[202] X. Zhang, X. Zhou, M. Lin, and J. Sun. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices, Dec. 2017.

[203] B. Zhao, Y. Tan, W.-Y. Tsai, J. Qi, C. Xie, L. Lu, and L. H. Schwartz. Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Scientific Reports*, 6(1):23428, Mar. 2016. ISSN 2045-2322. doi: 10.1038/srep23428.

[204] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong. Image BERT Pre-training with Online Tokenizer. In *International Conference on Learning Representations*, Oct. 2021.

[205] M. Zhu, B. Ren, R. Richards, M. Suriawinata, N. Tomita, and S. Hassanpour. Development and evaluation of a deep neural network for histologic classification of renal cell carcinoma on biopsy and surgical resection slides. *Scientific Reports*, 11 (1):7080, Mar. 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-86540-4.

[206] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le. Learning Transferable Architectures for Scalable Image Recognition, Apr. 2018.