# Interpretable AI Methods for Breast Cancer Whole Slide Image Analysis

**Thomas Harry Allcock**

Submitted in accordance with the requirements

for the degree of Doctor of Philosophy

The University of Leeds

School of Computing

May 2024

The candidate confirms that the work submitted is his/her own, except where work which has formed part of a jointly authored publication has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Some parts of the work presented in Chapters 6 published in:

Allcock, T., Bulpitt, A., Hanby, A., Millican–Slater, Cross-attention multiple instance learning for interpretable whole slide image classification. In 27th Conference on Medical Image Understanding and Analysis 2023 (p. 206).

The above publications are primarily the work of the candidate.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis maybe published without proper acknowledgement.
©2024 The University of Leeds and Thomas Harry Allcock

## Acknowledgements

First of all, I would like to express my gratitude to my research supervisors Professor Andy Bulpitt, Professor Andrew Hanby and Dr Rebecca Millican-Slater, whose knowledge and guidance have been invaluable to this project. Their advice has helped me to overcome difficult obstacles and they have constantly steered me in the right direction.

I would like to thank the Doctoral Training Program and the leads Professor David Hogg, Professor Vania Dimitrova and Mr Owen Johnson for their constant support and commitment and for all the opportunities they have provided their students. I also would like to thank the many students on the training program for all the occasions we have discussed ideas and exciting research directions.

This work was also supported by Leica Biosystems and I would like to extend my thanks to those there who helped and supported me with this project. This work would also not have been possible without the data provided by patients and collected by the NHS as part of their care and support at Leeds Teaching Hospital Trust.

Finally, I extend my heartfelt appreciation to my family and friends. To Rosie, thank you for your unwavering encouragement, understanding and moral support. To my parents Gill and Tony and my sister Sophie, I am forever grateful for your constant belief in me which has always motivated me during challenging times.

## Abstract

Histopathology is the diagnosis and study of diseases of the tissues which is undertaken by pathologists. For cancer diagnosis, they would traditionally examine tissue under the microscope and provide information to clinicians regarding cancer grade, type and potential response to treatment. However, with the introduction of digital pathology, a transition has begun away from traditional whole slide interpretation using microscopes. Instead, slides can now be viewed digitally, allowing for greater ease of use, rapid peer review and more efficient storage. Also, the availability of thousands of digitised histopathology slides facilitates the development of Artificial Intelligence based tools which can provide quick and accurate patient diagnostics. However, one drawback of these methods is, as their architectures have become more sophisticated, and performance increases, our ability to explain their decisions has decreased. The field of explainable AI attempts to remedy this by providing both models and tools for understanding the decision making process of AI solutions. In this thesis, methods for increasing the interpretability of Breast Cancer diagnostics models are explored. The main contributions are divided into three distinct chapters. In the first, a prototype based model, originally designed for fine-grained natural image classification, is used for interpretable breast cancer sub-typing and grading for the first time. The method achieves superior performance to less interpretable methods. The second contribution makes use of an attention based approach which allows for inspection of the most important image regions towards a classification. The approach is applied to a novel classification task which predicts the Nottingham Prognostic Index group. The approach is additionally extended to utilise both a patient's primary tumour site and lymph node image through late slide fusion, the latter of which is often neglected in automated diagnostic tools. The third contribution aims to combine the prototypical approach with the attention based approach. By making use of the prototype based approach, greater explainability can be provided to the attention mechanism. The methodology is used for both breast cancer grading and lung cancer sub-type prediction. Comparable performance is found with other state-of-the-art methods for lung cancer sub-typing and superior performance is found for breast cancer grading.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Overview

Breast Cancer is the most commonly diagnosed cancer and the leading cause of cancer death among females globally. Breast cancer in women accounted for almost 1 in 4 cancer cases among women worldwide in 2018 [19] and is the most common cancer in the UK, accounting for 15% of all new cancer cases (2016-2018) [21]. Early detection of cancers is crucial for patient survival and overall outcomes as early stage, less aggressive cancers are easier to treat. There is also more time to explore effective treatment options, including surgery, radiation therapy, and chemotherapy.

Detection of a possible breast cancer can be done clinically via breast examination and/or via radiological imaging modalities such as mammography, breast ultrasound or magnetic resonance imaging (MRI). Confirmation of the diagnosis is made via histopathological examination a of a tissue sample i.e. a biopsy. The latter is the gold standard for cancer diagnosis. Here tissue specimens are extracted from the body and examined under a microscope by a trained pathologist. The morphological appearance of the tissue is assessed by the pathologist to determine whether cancer is present within the tissue. Additionally, several scoring mechanisms have been developed through examination of large cohorts of tumours with outcome data which allow pathologists to assess the stage and aggressiveness of any present cancer, such as the Gleason score for prostate cancer, or the Nottingham prognostic index for breast cancer. Different types and stages of cancer require specific treatments. By diagnosing the type, location, and extent of cancer, pathologists can develop personalised treatment plans tailored to the individual's condition and increasing the chances of survival.

Due to the quantity of breast cancer cases each year there is a need for fast and accurate diagnosis. Unfortunately, the diagnostic process is time-consuming. From biopsy extraction and slide generation, to pathologist examination and treatment suggestion, there are many steps before treatments can begin. Additionally, although the efficacy of many cancer staging scores have been shown, there is still inter-pathologist variability [54], especially for borderline cases. This is often due to the subjective nature of some tasks such as nuclear pleomorphism scoring for breast cancer grading where a value for the difference in appearance between the tumour cells present and healthy cells is given. Other tasks such as mitotic counting also results in variability. Due to the small appearance and overall quantity of cells present this can be time-consuming and fatiguing and is therefore more susceptible to human error.

Computer-Assisted Diagnostic (CAD) tools have been proposed as one solution to faster and more accurate cancer diagnostics across a variety of medical imaging domains, including histopathology. In particular, artificial intelligence (AI) model based solutions [15] have emerged as a leading tool for complex image analysis. With the introduction of digital histopathology, tissue slides, which were traditionally analysed under a microscope, can now be viewed digitally after being imaged by a powerful, high magnification scanner. This has opened up many opportunities for CAD tools to analyse histopathology slides and assist with many diagnostic tasks such as tumour detection, cell counting and bio-marker prediction. Ultimately the goal of CAD systems is to improve workflow efficiency by streamlining the interpretation process, allowing for faster image analysis and report generation, ultimately leading to more timely diagnosis and treatment planning.

However, there are several challenges which must be addressed before AI based CAD tools can see widespread implementation into routine clinical practice. One important challenge is Interpretable and Explainable AI (XAI). Making sure AI solutions are accurate and can provide sufficient reasoning and explanations for predictions is extremely important, especially in a high risk environment such as healthcare. It is imperative that pathologists and clinicians understand the reasons behind a model prediction in order to better inform treatment plans and to avoid incorrect model predictions. These challenges are discussed in more depth in section 1.2.1

## 1.2 Research Aim and Objectives

The main aim of this work is to develop and evaluate computational models for automated histopathology slide analysis for Breast cancer diagnosis. This is an exciting and constantly evolving research space due to the possible benefits such methods could provide. From faster and more reliable diagnosis, to development of novel biomarker prediction, the inclusion of AI solutions in digital pathology is imperative. The focus will be on the development and validation of interpretable models in order to overcome the limitations of opaque models. Therefore, the main contributions and objectives of this work are as follows:

1. **The effectiveness of Prototypical Part Networks within histopathology is evaluated**. These are intrinsically interpretable Convolutional-Neural Network (CNN) based models which incorporate prototype layers for modern case based reasoning. They have been largely unexplored within the context of digital pathology but have shown success in other medical imaging modalities. The choice of pooling function and use of an additional, orthogonal loss term are found to be crucial for effective prediction of both breast cancer sub-type and grade from two open source datasets.

2. **A Novel Breast Cancer dataset is explored**. This dataset contains both a patient's breast primary tumour slide and a lymph node slide. The spread of cancer to lymph nodes is a sign of tumour growth and aggressiveness, however they are often overlooked in automated cancer diagnosis. This dataset allows for exploration of techniques which combine both slide types. The use of interpretable attention based multiple instance learning (MIL) approaches are explored along with several fusion operators for prediction of the Nottingham Prognostic Index (NPI).

3. **Cross-Attention Multiple Instance Learning is proposed** to provide greater interpretability to the original attention based MIL approach, and additionally refine the feature space using prototypes. This aims to incorporate some of the insights gained from

the Prototypical Part Network into the attention based MIL approach. The proposed approach additionally outperforms other state-of-the-art (SOTA) methods on breast cancer grading and lung sub-type prediction.

### 1.2.1 The Challenges of AI solutions in Histopathology

There are many challenges which must be overcome before AI solutions can be effectively implemented into the treatment pathway and pathologists can make use of the many benefits that they offer.

The sheer size and scale of pathology slides, also known as a whole slide image (WSI), is a major challenge for any CAD tool, as shown in Fig 1.1. Unlike other medical image modalities, such as MRI or computerised tomography (CT) scans, histopathology images are extremely large, often several giga-pixels in size. Because of this, they can contain a variety of different cellular and tissue structures, as shown in Fig 1.2. Each of these different structures can be individually classified along with the thousands of cells present in a slide. Therefore, one challenge is determining what information within a slide is relevant to a particular diagnostic task. Of course, one could classify and segment all possible morphologies making use of all the available data, however this would not only be time consuming and expensive, but the large number of features found does not guarantee optimal model performance, as it is likely that most are unrelated to the task at hand. Determining the level of detail required is also an important challenge. WSI are composed of multiple resolutions or magnifications. As the resolution increases the level of detail also increases, but so does the size of the image. At the highest magnifications the images are too big to feed directly into any AI based tool, such as a Deep Neural Network (DNN). Therefore, regions of interest (ROIs) must be cropped from the image in order to make use of the fine-grained details. However, in doing this much of the slide context can be lost. Alternatively, low magnification images can be used to increase the spatial context at the cost of fine-grained detail. Finding a balance between both views which is optimal for a specific medical task is a major challenge that needs to be addressed with any CAD tool.

Annotations are also a major challenge to address in the domain. Annotations provide

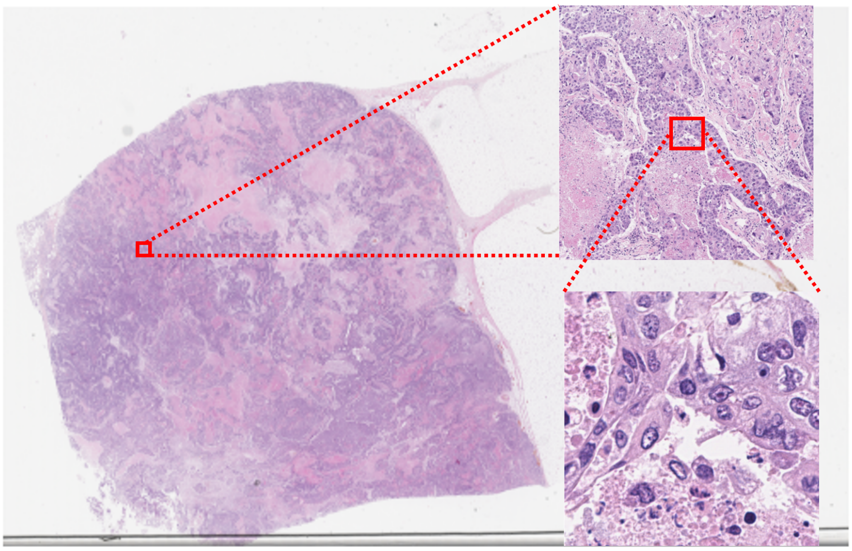


Figure 1.1: The scale of a pathology slide. Slide image is shown on the left with crops taken at different zooms on the right to show the scale and amount of detail contained in a slide.

training and testing labels, allowing for effective development of any diagnostic model. Labels provide a more obvious training approach and generally produce the best performance. They also allow for easier model evaluation. Annotations can be strong, allowing for prediction of individual cells, tissues or specific bio-markers. Alternatively, annotations can be weak where only a few labels are available per slide, or only a single prognostic factor or biomarker is known per patient. Dense annotated data is particular expensive to collect. This is because one, histopathology slides are extremely large so time-consuming to annotate, and two, because annotations require domain expertise making them very difficult to crowd-source. To overcome this challenge streamlined and effective methods for data collection and annotation need to be put in place [4], or effective methods for utilising as few annotations as possible need to be explored [77].

Model generalisability is another challenge. Every medical centre has different scanners and staining protocols which results in slightly different slide appearances. This is largely inconsequential for human observers, however automated analysis tools can be highly impacted by subtle changes from the source image domain. Often AI tools are developed and validated using data from a single institution and therefore do not translate well to other institutions

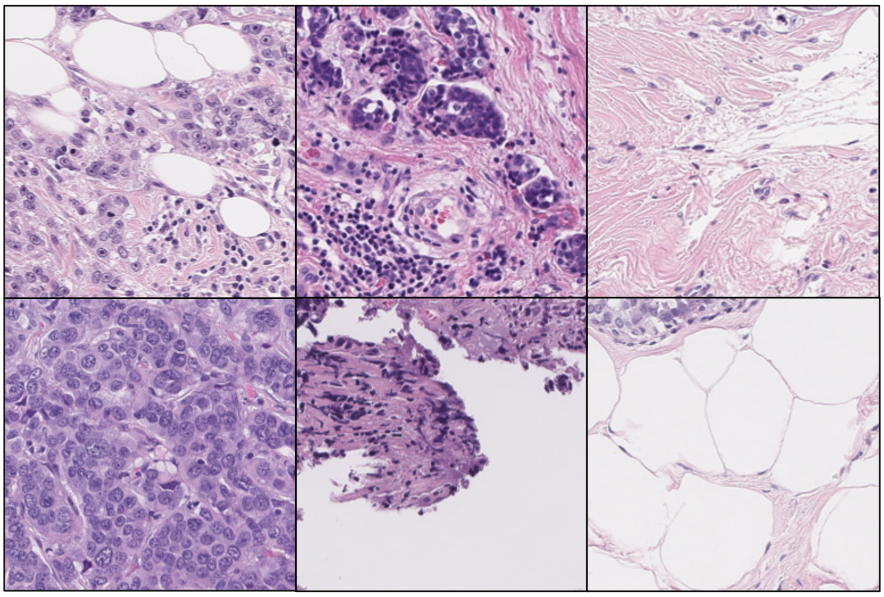


Figure 1.2: Tissue patches extracted from WSI to showcase the variety of morphologies present.

when implemented. Therefore, developed tools need to be trained and validated using slides from different institutions to assess their robustness [7]. Alternatively, methods for stain normalisation across domains could be developed [125] or computer vision tools could be designed to focus on the underlying biology, such are cell shapes and sizes, compared to the colour and textures within an image.

A final challenge, and one that is of particular interest to this work, is model explainability and interpretability.

### 1.2.2 The case for Explainable AI.

With the advent of deep learning, and the development of DNNs, predictive models have only become more complex. This increased complexity has resulted in SOTA performance in many computer vision benchmarks [75] and the adoption of DNNs into a wealth of domains. However, this increased complexity has brought with it a lack of interpretability. DNNs are often considered to embody a "black-box" [109] after they have been trained. The millions of mathematical processes which occur inside a network to produce an output means there is no way to verify the decision it makes. Of course, it can be stated through the combination of layers and neurons “how” a model works but “why” a model reached a certain decision is the far more beneficial question. In a high-stakes decision environment, it is crucial that models can be questioned and interpreted by pathologists. This will allow pathologists to: 1) Understand the reasons behind a diagnosis and overturn any prediction they feel is based upon sparse evidence. 2) Discover bias and inaccuracies within the model. This would allow diagnostic tools to be improved over time while being fit for clinical use. 3) Create greater trust between the pathologist and automated tool, thus increasing the uptake of tools into clinical practice. Furthermore, deep learning models developed on medical data are often trained without much inclusion of domain knowledge. Although they show improved performance they simultaneously reduce human involvement in knowledge distillation. Much information is hidden in arbitrarily high dimensional spaces which is unavailable for human questioning. As a result, XAI approaches could generate new hypotheses and provide means for human understanding [63].

### 1.2.3 The Explainability Issue.

The first obstacle to explainability is determining semantically meaningful interpretations of a model’s decision boundary. DNNs are trained on large sets of complicated data and learn abstract features as a result. However, these abstract features are not guaranteed to represent features interpretable to humans. If you could faithfully explain a DNNs decision it might not provide any further insight into the inner workings. Therefore, any explanation must be presented to align with the way humans think. The second obstacle is determining what counts as an explanation. Often this requires expertise in the domain of interest in order to identify the form of an explanation that is satisfactory. It is of course possible that too grand an explanation may result in an information overload and hinder the user experience of an automated system. This is all too important in digital pathology where pathologists require efficient insights rather than bulky explanations. To help resolve this issue, several works have attempted to create a theoretical framework for what constitutes as an explanation [99, 101]. The final obstacle is producing unbiased model explanations. There is of course a strong incentive to promote trust in the AI predictions, however explanation designs could run the risk of being more persuasive than informative, ultimately misleading the user [111].

## 1.3 Thesis Structure

A brief summary of the following chapters will now be introduced to provide an overview of the Thesis. This thesis is divided into 6 subsequent chapters. These are the background, dataset descriptions, three research chapters each highlighting a particular contribution and finally a conclusions chapter.

- **Chapter 2:** Here most of the background for the work developed is introduced. Initially, the origins of cancer are introduced along with specifics on breast cancer and its various types. Common prognostic factors will be explained such as the grading and NPI. Following this, methods for automated cancer diagnosis will be described, from early hand-crafted features to new, neural network based approaches such as CNNs and transformers. XAI techniques for computer vision are also presented along with their uses in computational pathology. Finally, the various performance metrics used throughout this work are described.

- **Chapter 3:** The four datasets used throughout this work are introduced. These datasets are the BACH dataset for breast cancer sub-typing, the breast cancer Grading dataset, the Leeds Teaching Hospital Trust (LTHT) dataset containing patient primary tumour sites and lymph node images, and The Cancer Genome Atlas (TCGA) lung sub-typing dataset. The curation of a novel breast cancer dataset is described in this chapter. The process of WSI segmentation, preparation, patching and feature extraction are also discussed. The way slides are partitioned into development and test sets is not included here, instead this information can be found in the specific chapters where these datasets are used.

- **Chapter 4:** The Prototypical Part Network [26], originally designed for interpretable fine-grained classification is introduced into the context of digital pathology. The model is evaluated on two ROI classification tasks. These are breast cancer grading and sub-typing. ResNet-18 and VGG-11 backbones are tested along with several options for pooling operators, loss functions and similarity metrics.

- **Chapter 5:** Attention based MIL approaches are explored for NPI group prediction. This chapter uses the LTHT data introduced in Chapter 3. Two approaches are explored. The first predicts the individual components of the NPI to determine the group. The second approach directly predicts the NPI group but using several fusion operators to fuse the attention head outputs between patient slides. The individual prediction approach is found to be superior, however the direct approach could be utilised for future tasks, such as survival prediction.

- **Chapter 6:** Cross-Attention MIL is introduced which proposes a prototype approach to the problem. This chapter additionally uses the data introduced in Chapter 3. The Cross-Attention approach aims to disentangle a WSI into the core tissue components and rank their importance towards a particular classification.

- **Chapter 7:** A summary of the work presented here is provided along with what future work could be conducted as well as limitations with the discussed methods.

# Chapter 2

# Background

## 2.1 Cancer

Cancer is the result of uncontrolled, and continual reproduction, or proliferation of cells. Cancerous cells do not respond appropriately to all signals within the body that control and regulate normal cell behaviour. Instead, cancerous cells will divide uncontrollably and eventually invade normal tissue and distant organs resulting in cancer spread throughout the body, known as metastasis [37]. A major distinction between cancer cells and normal cells is the display of density-dependent inhibition of cell proliferation. For normal cells, growth is limited once cells reach a certain cell density. Cancer cells alternatively are not sensitive to the current cell density and continue growing to extreme cell densities. The initial development of cancer cells is due to genetic mutations from environmental and lifestyle factors, as well as inherited genetics.

Within cancer pathology there exist both benign and malignant tumours. A tumour can be thought of as a mass of abnormal cells. Benign tumours are largely non life threatening and are confined to their location of origin. Malignant tumours, on the other hand, are much more dangerous as they have properties allowing them to spread to distant regions within the body. This allows them to interfere with regular bodily functions and prevent healthy cells from conducting their normal behaviour. Therefore, effectively distinguishing between these two tumour types is crucial.

Cancers can generally belong to one of three different groups. The majority of cancers are Carcinomas which arise from the epithelial cells of healthy tissue. Sarcomas on the other hand are much rarer conditions and are found within the connective tissue of the body. This could be in muscles or bones for example. Finally, Leukemias and Lymphomas arise from the blood-forming and immune system cells. Additionally, tumours can be classified based on their site of origin, for example, lung or breast cancers which come with their own levels of severity and treatment pathways.

## 2.2 Breast Cancer

Breast cancer refers to cancer which originates from breast tissue, and most commonly from the inner lining of milk ducts or the lobules that supply the ducts with milk. In general there are two main types of tissue in the breast. These are the glandular tissue and stromal tissue. The glandular tissues contain the milk-producing glands known as lobules as well as the ducts which are the milk passages. The stromal tissue contains other tissue types such as fatty or fibrous connective tissue.

Historically it was believed that most breast cancers begin in the cells that line the ducts, known as ductal carcinoma and some cancers arising from the lining of the lobules, known as lobular carcinomas. Over the last thirty years the taxonomy of breast cancer has been shown to be far more complex. Nearly all mammary carcinomas are believed to derive from stem cells located in the terminal duct lobular unit (TDLU) [153]. The diversity of morphology derives from the molecular pathways along which carcinomas evolve. In some cases fairly specific molecular events determine morphology.

Tumours with fairly distinct morphological appearances are called “*special-type carcinoma*” of which lobular carcinoma is one example. Additionally, there remain a fairly large group of tumours which do not have such an easily reproducible morphology and these are put in a heterogeneous group called ductal carcinoma of no special type (NST). In reality, the taxonomy of breast cancer has become increasingly more diverse and complex and the details are beyond the scope of this work [43].

### 2.2.1 Types of Breast Cancer

**Non-Invasive Breast Cancer**

Also referred to as in-situ carcinoma, cancer cells are confined within a ductal or lobular structure and are currently not invading surrounding connective breast tissue. Ductal carcinoma in-situ (DCIS) is the most commom form of non-invasive breast cancer. Lobular carcinoma in-situ (LCIS) can occur and is treated as a marker for increased cancer risk and longitudinal studies show that the rate of progression to invasion is extremely certain. In-situ carcinomas are an earlier stage in the development of malignancy in which the disease is confined to the microanotomical structure from which it has arisen. All invasive breast cancers will have a precursor in situ phase, which can be either long or short term. It should be noted that some low grade DCIS share similar genetic features to LCIS leading to studies to determine whether it may behave in the same manner.

**Invasive Breast Cancer**

This type of cancer is when cancerous cells are no longer confined and instead break through the ductal or lobular boundary, invading the surrounding stromal tissue. Invasive ductal carcinoma is the most common type of breast cancer and is identified in around 80% of all breast cancer diagnostic cases [3]. Invasive breast cancer is considered more severe than non-invasive carcinoma.

### 2.2.2 Sub-typing

Due to the heterogeneous nature of breast cancer there are also various sub-types which can impact which treatments are available to patients. These sub-types are commonly grouped into four categories based on the immunohistochemical expression of hormone receptors. Receptors are proteins embedded in the call membrane and bind to specific substances in the blood. Breast cells can have receptors for estrogen or progesterone which are required for the cell to grow. Immunohistochemistry (IHC) is used to detect the presence of specific protein markers which makes use of the anti-gen anti-body interaction to identify the presence of these proteins. These four sub-types are: estrogen receptor positive (ER+), progesterone receptor positive (PR+), human epidermal growth factor receptor positive (HER2+), and triple-negative (TNBC) which is ER-negative, PR-negative, and HER2-negative. Each sub-type is determined from the expression of the specific receptors [107]. Additionally, expression of the Ki67 antigen is an excellent marker for providing information on cell proliferation. High expressions of Ki67 are associated with more aggressive tumour proliferation [85].

#### Luminal-A

Luminal-A tumours are characterised by the presence of ER and/or PR and the absence of HER2. They are generally low grade and grow slowly. Because of this they have the best prognosis and are associated with a higher survival rate and lower chance of relapse. Luminal-A also have a lower expression of Ki67.

#### Luminal-B

Luminal-B tumours also show expression for ER and do not necessarily express PR however they show a greater expression of Ki67. This sub-type has a worse prognosis compared to Luminal-A and are generally of intermediate to higher grade. They benefit from hormone therapy and a higher percentage from chemotherapy compared to the previous group.

#### HER2

This sub-type shows an over-expression for HER2 with no expression of ER and PR. HER2 sub-types have a worse prognosis compared to their luminal counterparts. They also require specific HER2-targeted therapies. These treatments can be very effective, meaning that the prognosis for HER2 has improved after their introduction.

#### Triple-Negative Breast Cancer (TNBC)

This is negative for all three receptors and is highly aggressive with an advanced stage and higher histologic grade which are discussed later in sections 2.2.3 and 2.2.4. They have characteristics such as; high mitotic count, scant stromal content, central necrosis and a stromal lymphocytic response. TNBC is poorly differentiated, meaning the cells look and behave very differently to normal cells. TNBC cannot be targeted with hormone therapy as the cancer cells don't have hormone receptors.

### 2.2.3 Staging

Breast Cancer Staging provides a measure for how much the cancer has spread around the body. Lower stages classify cancers which are still within their site of origin while much higher stages classify cancers with distant metastases. Specifically, breast cancer staging is performed using the **TNM** Staging scheme which is divided into three scores: **T** describes the size of the primary tumour, **N** describes if cancer spread has occurred within nearby lymph nodes and to what extent, **M** indicates whether cancer has metastasised to distant organs [133].

### 2.2.4 Grading

Breast Cancer Grading is a tool to determine how likely and fast breast cancer is to spread. Although staging allows us to understand the current extent of a cancer, the grading can determine the degree of potential malignancy. The Bloom-Richardson Grading scheme [17] is the most used grading scheme for breast cancer and groups patients into Grade 1 (Low), Grade 2 (Moderate) and Grade 3 (High) grades depending on three factors: Tubule formation, Nuclear pleomorphism and mitotic count. For each of the three factors a score of 1-3 is given. A total score of 3-5 gives Grade 1, a score of 6-7 gives Grade 2 and a score of 8-9 gives Grade 3.

#### Tubule Formation

The first factor determines the extent to which cells within the duct or lobule are still forming tubular structures, which are round shapes around a central space. A low score is given for cases which regularly form these structures while a higher score is given to sections where cells are arranged in a sheet like formation.

#### Nuclear Pleomorphism

The second factor assesses how much cancerous cells resemble healthy, normal epithelial cells within the breast. Lower grades have uniform cells with small nuclei and are a similar size to healthy epithelial cells. High grades have irregularly shaped cells with enlarged nucleoli. These pleomorphic changes are a sign of abnormal cell reproduction.

#### Mitotic Count

The third factor counts the number of mitotic figures across 10 consecutive high power fields (HPF) of the most mitotically active area across a H&E stained WSI [129]. One HPF is the area seen with the 40x objective. Mitotic figures are cells which are in the process of dividing. Therefore, larger quantities of dividing cells indicate a more rapidly growing cancer and suggest a higher score to contribute to the grade.

### 2.2.5 Nottingham Prognostic Index

The Nottingham Prognostic Index (NPI) is a prognostic marker developed for breast cancer prognosis. NPI can help to inform the decision of whether or not to undergo adjuvant systematic treatments, such as chemotherapy following surgery. This is because the benefits of chemotherapy are proportional to the risk of death from death cancer. Therefore cases with a higher NPI score would benefit more from such treatments. Additionally, NPI can be used for counselling purposes where a patient can be given the risks associated with their tumour [78].

NPI can assist in stratifying patients into 6 groups based on their 10 year survival. These groups can be seen in Table 2.1. NPI was introduced by Galea in 1982 [53] and used to predict 10 year survival of 80%, 42% and 13% in three groups of patients. They found that, although a number of factors were related to survival in univariate analysis on 387 patients, there were only three factors that were also significant on multivariate analysis. These factors were the maximum diameter of the tumour, the histologic grade and finally the lymph node stage. The index is therefore determined using a linear combination of these three important prognostic factors shown in equation 2.1.

$$NPI = (0.2 \times TumourSize) + Grade + LymphNodeStage \tag{2.1}$$

The tumour size is measured in centimetres and is based on the size of the greatest dimension of invasive cancer in histological sections. In multifocal breast cancer, which is the result of a

Table 2.1: Prognostic groups from NPI

| **Prognostic Group** | **NPI** | **10 year survival (%)** |
|---|---|---|
| Excellent | $< 2.4$ | 96 |
| Good | 2.4 - <3.4 | 93 |
| Moderate-1 | 3.4 - <4.4 | 81 |
| Moderate-2 | 4.4 - <5.4 | 74 |
| Poor | 5.4 - <6.4 | 55 |
| Very Poor | $\geq$6.4 | 38 |

single tumour cell clone spreading and developing disease independently at various locations, the largest tumour mass is used. Greater tumour size is associated with poorer prognosis [23]. The lymph node stage is particularly informative for breast cancer analysis as greater nodal involvement leads to much worse prognosis [23]. This component takes on a score of 1, 2 or 3 dependent on the extent to which nodes are involved. A score of 1 is given to patients whose nodes have no involvement at all. A score of 2 is given if either, up to 3 axillary nodes or if the internal mammary node alone is involved. Finally, a score of 3 is given if there are at least 4 positive axillary nodes or any low axillary node and an internal mammary node are involved together. Axillary nodes are nodes found in the human armpit, while low axillary nodes are found in the lower armpit. Most of the lymph drained from the breast passes to the axillary nodes making them an initial indicator for metastatic tissue. Internal mammary nodes lie along each side of the breast bone and are also involved in the drainage of lymph from the breast. The score for grade is the same as described in the section 2.2.4 above.

Overall, despite the in-depth molecular complexity of breast cancers, at a practical level marker status permits division into Luminal A, B, HER2 and TNBC, which combined with grade, NPI and stage, provides a useful road map for the day to day treatment of breast cancers.

## 2.3 Computational Pathology

With the introduction of digital pathology, new possibilities for computer-assisted analysis of WSIs have opened up, which can be summarised in Fig 2.1. The earliest works made use of hand-crafted features. These are hand-engineered features developed predominantly for a particular use case and are not applicable for a broad use. Later works made use of deep learning approaches to model the underling data distribution. Such methods are extremely powerful but can be opaque and lack the biologically aligned features of hand-crafted approaches. More recent methods have aimed to incorporate powerful transformer and graph-based [2] methods into the pipeline. These methods can model the interactions between cells or tissue regions as well as their spatial arrangements. Ultimately, a variety of computational methods can be applied to digital images to assist pathologists in tasks such as automated cell counting, tumour identification and quantification of biomarkers. In this section, traditional methodologies for WSI analysis will be reviewed as well as the current SOTA.

Computational Pathology can be broken down into two broad approaches. The first encompasses supervised learning techniques. These techniques learn a mapping from inputs to outputs so that predictions can be made on unseen data. This leverages large quantities of annotated data to train the computational models. In computational pathology the annotated data is often in the form of tissue pixel maps. These highlight the regions of important tissue within a WSI. For example, regions of tumour and normal tissue could be annotated to train a tumour classification model. Other annotations could include outlines of cells for cell detection as well as labels for each cell indicating its type (e.g lymphocyte or tumour cell). Because of these provided annotations, models trained through this approach often have the highest performance and methods can be easily validated against the ground-truth. One downside to supervised schemes is the requirement for large sets of annotated data. Especially in the area of histopathology, annotations are time-consuming and expensive to produce due to both the scale of WSI and the need for domain expertise. Unlike the situation with natural datasets, crowd-sourcing is much harder for histopathology datasets as annotators must have medical training to provide accurate labelled data.

The second approach uses weakly-supervised/unsupervised learning techniques. These methods train models using limited label data or even no labelled data at all. When there are no labels available a model aims to learn patterns from unlabelled data meaning it is not provided with explicit input-output pairs. For weakly-supervised learning the labels provided for training are less informative or less granular than the precise labels provided in fully supervised learning. For example, the annotated data used could be a small set of finely annotated images, point annotations which use a single point as the label for certain tissues thus any tissue close to that point is considered to have that label or finally just a WSI label, such as the tumour stage or another biomarker. For unsupervised learning, clustering is a very common technique which aims to classify different cell and tissue types by clustering their representations in the feature space. For example, cells could be clustered together because they have similar lengths or colours. The most common approach to weakly supervised learning is MIL [46] which often uses just the WSI label for training. The benefit of weakly and unsupervised approaches is that finely-annotated large scale datasets are not a requirement, allowing for easier and quicker model development. They also allow for analysis of novel biomarkers where fine grained annotations are unavailable. However, these methods on average perform worse than fully-supervised approaches as there are no precise labels to guide training. This is particularly true for precise tasks such as tumour and cell segmentation [13]. In some instance more WSI can be required as the training signal is weaker because of noisier and less precise labels. These less precise labels come about because there can be many features within a WSI that do not correspond to the global slide label. For example, a slide containing tumour would be given a malignant label however there is likely plenty of healthy tissue still present on the slide, which makes the label less precise. On the other hand, the ability to learn from less densely labelled data can allow for very large datasets to be used which in turn can increase model performance compared to supervised methods on some tasks. MIL methodologies are used heavily in both chapter 5 and 6 of this work.

Before discussing the approaches in depth it is important to understand the standard pipeline for WSI analysis as the majority of approaches in this work and many others make use of the same pipeline. Firstly, a digital slide needs to be pre-processed into a more user friendly format. Digitised histopathology slides can be several giga-pixels in size making it infeasible to process as a whole using common image processing and neural networks approaches due to memory

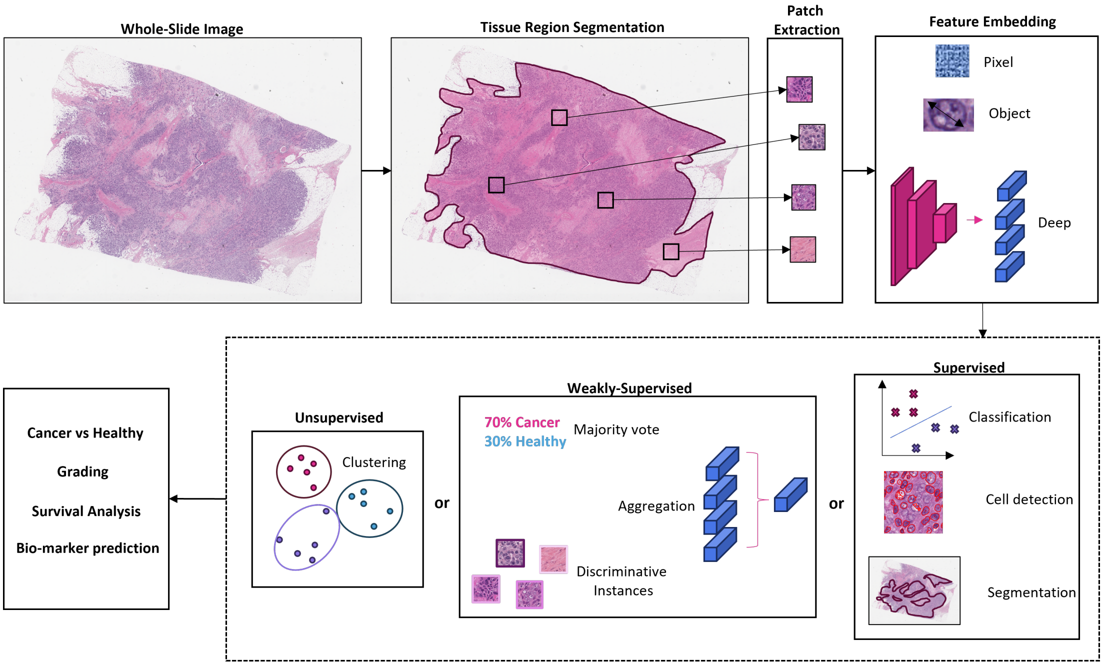


Figure 2.1: Overview of computational methodologies used for Whole-Slide Image analysis.

limitations. The most common approach for dealing with this is splitting a WSI into patches using a sliding window approach. Patches are generally extracted from all non-background regions and are non-overlapping, meaning that each region within the slide only appears once among all patches. Overlapping regions can be used when aiming for more refined classification or segmentation approaches. Alternatively, patches can be extracted from important slide regions only, thus reducing the computational complexity. This can be done with annotations provided by pathologists or by using an attention mechanism to crop important patches [160]. Following the extraction of patches across a slide, they can now be classified into different tissue categories, segmented for cells and tissue regions or passed through a pre-trained CNN, such as a ResNet to extract deep features from the image. If the final goal was to classify patches or perform segmentation then often the process stops here. However, in cases where we would like to predict a particular biomarker, such as the tumour-stroma ratio across a slide or perhaps predict patient survival, the information extracted across all patch slides needs to be aggregated together. This can be achieved in a variety of ways which will be explored in more depth later.

## 2.4 Hand-crafted features

For the task of patch prediction, initial methodologies focused on the use of hand-crafted features [73]. Hand-crafted features, or content-based features refer to manually designed image representations which aim to capture specific semantic information which can be used to categorise or increase the understanding of a particular image. The process of designing such features can require specific domain knowledge and expertise depending on the feature. Relevant visual patterns need to be observed within an image and algorithms need to be designed to extract those patterns from it. Within the context of histopathology image analysis there are three main types of features: pixel, object and semantic-level features. Pixel-level features focus on raw image information such as image pixels as well as primitive image features such as colour and texture. Object features build upon pixel features to segment images into some meaningful regions or objects. Semantic features place objects and regions identified in the context of whole image scenes. Therefore, as features go from the pixel-level to the semantic-level the amount of raw data they capture decreases in favour of carefully designed high-level image features. These higher level image features often have greater biological interpretability but require domain expertise to create[170].

### Pixel features

Pixel level features capture properties such as colour and texture from image pixels and are the least interpretable form of hand-crafted feature. Colour features can be are extracted from different colour spaces such as red-green-blue (RBG), hue-saturation-value (HSV) or CIELAB. Textural features on the other hand look for contrast, sharpness, intensity and edges using grey-level intensity profiles, Haralick grey-level co-occurrence matrix (GLCM) features or wavelet and multiwavelet submatrices. Despite lacking biological interpretability, they have been successfully used for many data-driven models, for example colour texture features extracted from GLCM were used for follicular lymphoma grading [119].

### Object features

Object level features describe properties of structures within histopathology images, such as cells, nuclei and glands. To collect object-based features, structures first need to be segmented. The quality of the segmentation greatly impacts the features extracted from the objects and is therefore an important step. Contour-based features look at the properties of an object boundary. This includes the perimeter, boundary fractal dimension or bending energy. Additionally, they include coefficients of parametric shape models such as Fourier shape descriptors and elliptical models. Region-based features collect information such as object area or solidity. Topological features are also useful as they allow for the spatial distributions of cellular structures to be captured. Deluanay triangulations, Voronoi diagrams, and minimum spanning trees, are powerful tools for extracting topological features. Properties of graphs which are useful are lengths of edges, connectedness and compactness. The final types of object features look at object density or average distance between objects in a neighbourhood. Object features have been used for many histopathology analysis tasks, one particular use has been for analysis of features of the basal cell nuclei to separate malignant regions [74].

### Semantic features

Semantic level features boast the highest interpretability of all hand-crafted features but require the most domain expertise as well as large amounts of annotated data, making it the most expensive approach. Semantic level features are very specific to histopathology as they extract biological features only found within this imaging domain. Semantic features look at the particular properties of specific structures such as nuclei, necrosis, tumour cells, lymphocytes and much more [164]. Often the presence of these structural elements is measured, as well as their density or quantity. Alternatively, the ratio and co-occurrence of each of the elements can be found. Classification of the objects is often required first which involves segmentation as well as extraction of object features to perform the classification.

#### Limitations

Although hand-crafted features have shown promising results, they can be time-consuming and expensive to produce. Creation of features requires extensive evaluation and they often do not generalise well to new tasks. Additionally, domain knowledge is a barrier to entry as features need to be designed with the underlying histology in mind. To overcome these limitations, deep learning methods have appeared which are powerful data driven approaches to image analysis [136]. Due to the limitation of hand-crafted features and the benefits that deep learning methods provide the work presented in this thesis focuses on the application of deep learning methods.

## 2.5 Deep Learning Methods

### 2.5.1 Supervised Methods

Within the category of supervised methods classification and object detection/segmentation based models exist. Classification models aim to classify whole patches within a WSI. Patches could be classified as containing a particular tissue type or for the presence or absence of tumour. Object detection and segmentation deal with object position prediction, such as for for cell detection tasks, and then solve semantic segmentation or instance segmentation problems. This work mainly focuses on understanding WSIs at the patch level and is not concerned with cell detection and segmentation. Therefore, works related to classification tasks will be covered in more detail, but brief information on segmentation methods is also presented.

#### Classification

For the task of patch detection, the first works to step away from the use of hand-crafted features used shallow CNNs. Cruz-Roa et al. [41] used a simple 3-layer CNN for the identification of invasive ductal carcinoma in breast cancer images. This method outperformed the handcrafted methods by 5%. Litjens et al. [91] utilised CNNs for micro and macro metastasis detection in sentinel lymph nodes and also for prostate cancer detection. Bejnordi et al. [51] also made use of CNN models. Specifically three separate CNNs were used. The first model predicted for patches of epithelium, stroma and fat regions within a slide. The second model used the stroma patches to identify tumour-associated stroma. Finally the last model used 8 representative patches from the tumour-associated regions for WSI classification into invasive carcinoma or Normal/Benign cases. The final model used for WSI classification also out-performed their previous work which used handcrafted features extracted from the epithelium, stroma and fat slide regions. These features were input into a random forest classifier.

One disadvantage of the proposed methods is the long computational time required to carry out dense patch level prediction. Therefore, Cruz-Roa et al. [42] proposed High-throughput adaptive sampling for Whole-slide histopathology image analysis (HASHI). The method adaptively chooses slide regions where there is a high uncertainty of the tissue patch being invasive carcinoma or not. The idea is that regions of high uncertainty will require greater sampling, while regions of high confidence need less sampling, thus reducing the amount of redundant patches being processed. Attention mechanisms have also been explored to reduce the amount of patches processed by determining the most discriminative image regions. Qaiser et al. [112] apply an attention mechanism for the prediction of HER2 score. They use a Recurrent Neural network (RNN) to predict the next location in an image to look at, depending on what the current region of the image being looked at is. After $T$ iterations the model predicts the final HER2 score. Similarly, BenTaieb et al. [14] also used a RNN methodology to predict whether a lymph node slide contains metastatic tissue. Xu et al. [160] proposed a hybrid attention method to classify ROIs from pathology slides. A Hard-attention mechanism first crops a patch from the image and a second soft-attention model finds smaller patches within the larger one. A CNN extracts features from the smaller patches and combines this with the larger patches' location information. A long short-term memory network (LSTM), which is a variety of RNN capable of learning long-term dependencies, outputs the ROI classification at the time step and the next location to consider next.

So far each method has only considered the information contained within the patch and not the information from surrounding patches. The spatial location of a patch within a WSI can be very informative and improve model performance. Kong et al. [72] introduced Spatio-Net which integrates the CNN with a 2D-LSTM and updates the features of a patch depending on the 8 neighbouring patches surrounding it. Bejnordi et al. [12] encode patch context by feeding a much larger patch to the model at test time. A first model is trained with small patches of size 224 × 224 pixels and then a second CNN model is stacked on top of the first which then accepts patches of size 768 × 768 pixels. During this second training phase the parameters from the first model are frozen and only those from the second model are updated. Neural conditional random fields have also been proposed to consider the spatial correlations between neighbouring patches [88]. The work by Roy et al. [115] utilised auto-encoders trained with images at different magnifications, producing an embedding from the fusion of the different magnification outputs. This embedding was then used for tumour classification. The use of auto-encoders additionally helped them to enrich the feature space.

Some works have also combined CNNs with handcrafted features in order to enrich the feature space. He et al. [61] merge both handcrafted features and CNN features together before passing to a support vector machine (SVM) for classification. They also make use of transfer learning by using both an AlexNet and GoogleNet pretrained on ImageNet.

#### Segmentation and Cell Detection

Segmentation methods involve the classification of individual image pixels into particular classes. Cell detection is a difficult task due to the irregular appearance of cells and the fact that cell borders can touch and overlap making it hard to determine a boundary between objects. Early deep learning works utilised fully-convolutional network (FCN) based regression models for detecting cells [27] which use a downsampling convolutional branch followed by an upsampling branch with deconvolutional layers. To improve detection some works modify the loss function. Xie et al. [157] make use of a proximity patch which determines the distance between each pixel and the nearest human annotation. Then they modify the loss function, not to use a fixed weight for every training sample, but instead to allow the model to determine the weighting based on the mean value of the training proximity patch. This addresses the issue where there are more unannotated pixels in an image. Xing et al. [158] also make use of a proximity map but additionally use an ROI extraction task. To address the problem of touching cells Naylor et al. [105] use a distance map which records the distance between each cell pixel and each background pixel. Therefore, pixels closer to the centre of a cell would have the largest distance and the highest probability of belonging to a specific object. Graham et al. [56] proposed a unified FCN for instance segmentation and classification. This method additionally encodes the horizontal and vertical distance information of cell pixels to their centre of mass.

### 2.5.2 Weakly-Supervised Methods

Weakly-supervised methods aim to reduce the annotation burden placed on pathologists by providing methods that make use of limited information. Weakly supervised approaches make use of a small set of finely annotated images or point annotations which use a single point as the label for certain tissues. However, the weakly-supervised approaches covered here make use of a single global label for a WSI or region of the image, because these approaches are the most relevant to the work in this thesis.

Weakly-supervised approaches which make use of a global slide label are almost always formulated as a MIL problem. In MIL, the input to a model is called a bag and the aim is to assign each bag to a label $Y$. A bag is made up of multiple instances $X = \{x_1, ..., x_K\}$ and $K$ can vary in size for each bag in the dataset. As only the bag label in known, in the case of binary classification, each instance in the bag is assigned the label of the bag $y \in \{0, 1\}$. In the simplest form, the label of a bag $X$ is given by

$$c(X) = \begin{cases} 0, & \text{iff } \sum y_i = 0. \\ 1, & \text{otherwise.} \end{cases} \tag{2.2}$$

This is the standard multi-instance assumption and states that for a binary classification problem, a bag is positive if there exists at least one positive instance in the bag. MIL then uses a transformation $f$ and permutation-invariant transformation $g$ to predict the bag label.

$$c(X) = g(f(x_1), ..., f(x_K)) \tag{2.3}$$

This could be approached in two separate ways. One approach is to have $f$ be an instance classifier to predict the label of each instance, and $g$ be a pooling function such as max pooling that aggregates the instance scores to produce a bag label. Alternatively, $f$ can be a feature extractor and $g$ can aggregate the features into a bag level feature embedding which is then used to produce the bag label. The latter approach generally achieves superior performance due to the direct supervision of the bag embedding [150].

For WSI analysis the slide is considered the bag and patches extracted across non-background slide regions are considered the multiple instances. Due to the size of a WSI, bags can become very large and often contain noise from uninformative patches, which makes training difficult. To overcome this, many methods attempt to determine which instances are discriminative. Hou et al. [64] used an expectation maximisation based method which assumes that for each patch there is a hidden variable indicating whether a patch is discriminative or not. Discriminative patches are then classified and a class histogram of the patch-level predictions is the input to a linear multi-class logistic regression model or an SVM with Radial Basis Function (RBF) kernel. They argue that, because a WSI can contain hundreds of patches, the class histogram is robust to miss-classified patches. Alternatively, Zhao et al. [174] incorporate a feature selection method based on a histogram of each feature and then use the maximum mean discrepancy of the feature between positive and negative bags to evaluate its significance. Courtiol et al. [38] proposed a method known as CHOWDER which first uses a ResNet-50 pretrained on ImageNet to extract and store features from each patch in a WSI. Then one-dimensional convolutional layers return a single score for each feature embedding. Finally the top and bottom $R$ embedding scores are retained and passed through an MLP for slide classification. To provide a richer aggregation of slide feature embeddings Campanella et al. [20] incorporated an RNN which looks at the $R$ most interesting tiles within the slide in terms of the positive class probability. This approach achieved impressive AUC scores: 0.991 for prostate cancer detection, 0.966 for breast metastases detection and 0.991 for basal cell carcinoma detection. Other approaches aiming to improve the aggregation of patch embeddings often make use of an attention mechanism, such as that introduced by Ilse et al. [67]. This method produces an attention coefficient for each patch embedding by passing them through a neural network, then a weighted average over all patch embeddings is performed is produce the final slide feature embedding. Important instances will have greater representation in the final embedding which is then used for slide classification. CLAM [98] added a clustering methodology to refine the latent space. High attention patches are given the positive slide label and low attention patches are given the negative slide label. These pseudo labelled patches are also used to train linear layers which helps to refine the feature space. CLAM achieved an average AUC across a 10-fold validation split of $0.940 \pm 0.015$ for the detection of lymph node metastasis. Attention based MIL approaches have been used for a variety of histopathology tasks [140, 81, 83, 59, 97] including predicting cancer of unknown primary origin. Zhang et al. [171] utilised an attention approach as a patch recommendation system which finds highly important regions first, which are then used for the final classification of ROIs from WSI. Dual-stream MIL (DS-MIL) [79] introduced a novel MIL aggregator which models the relations between all patches and a critical instance. This critical instance is determined by using a classifier on each patch embedding which produces a score for how much the patch resembles the positive class. Max-pooling over these scores then determines the critical instance. The critical instance of a slide is therefore the patch which is the most likely to be present in a positive class. The use of a teacher student framework has also been explored for MIL [126] where initially a teacher model is trained using a max pooling MIL approach and patches with high confidence are used to train a student model. The student model thus produces more informative features which are then used for the final WSI classification. The work by Qu et al. [113] recognised that most methods only learn a bag-level or instance-level decision boundary and therefore do not model the data distribution. Instead, a cluster-conditioned feature distribution modelling method is proposed which first clusters the patch features from positive slides using K-means. Using Mahalanobis distance, a score measuring the minimum distance from all instances in positive and negative bags to the clusters is found. Higher scores indicate a higher probability the instance is positive.

Most MIL approaches applied to WSI first pre-process the patches using a pretrained feature extractor to reduce the memory consumption of their model. However, this prevents feature extraction being trained as the approach is not end-to-end. To resolve this Xie et al.[156] proposed a part learning approach where patches are mapped to one of $k$ global centroids and these slide specific centroids are then used to represent the $k$ parts of the slide. Sharma et al. [123] also introduced an end-to-end approach where patches from a slide are clustered into groups using features from a frozen feature extractor. Equal patches are then clustered from each group and passed through the online feature extractor and attention based MIL aggregator. The use of KL-divergence encourages patches from the same cluster to have similar attention values.

MIL approaches often require hundreds of WSI to produce robust results because the training signal can be weak due to the noisy labels. To combat this, Wang et al. [149] utilised a small sample of coarse annotations to guide an initial CNN to learn discriminative patches for each class. Features from these discriminative patches were then combined to produce per class embeddings which were later used for the final slide prediction. Limited annotations alongside an attention aggregator were also used for gastric cancer image classification [148]. Alternatively, Zhang et al. [169] split the initial bag into a set of pseudo bags by random patch selection; a two stage process which first produces a pseudo bag embedding and then combines these to give a final slide label. The use of pseudo bags is a form of augmentation helping to learn from fewer slides.

### Self-Supervised Learning

In most MIL approaches, the feature extraction stage is offline, meaning that the feature extraction uses a network with frozen parameters and patch features are then stored before training. This means, the feature extractor is not updated during training. Although, as previously mentioned, some works aim to allow for end-to-end training, other methods have looked at ways to improve the feature extractor first. This often involves an initial stage before feature extraction occurs where the feature extractor is pretrained to learn histopathology specific features. Unlike using natural image features, learning from histopathology images allows features to be tailored to the domain and is hypothesised to increase performance. These forms of approaches are called self-supervised methods. Models are trained initially through a specific task which does not require labelled data. For example, input images could be rotated from their original frame and the model must predict the degree of rotation. This form of training forces the model to understand the semantic information contained within the images in order to complete the task. The learned features can then be applied to downstream tasks and fine-tuned for particular classification problems using labelled data. The most common form of self-supervised learning in histopathology is contrastive learning. This approach aims to keep similar images close together in the latent space, while pushing apart dissimilar images. Dehaene et al. [44] used a contrastive learning approach known as MocoV2 [33]. They compared the patch features from ImageNet to those obtained after self-supervised pre-training and found that the Chowder method achieved a higher test AUC with the latter approach. Ciga et al. [36] found a similar result, but this time using the Sim-CLR [32] method.

### 2.5.3 Transformers

Transformers are a type of model architecture with applications in both supervised classification and segmentation tasks as well as weakly-supervised MIL frameworks. Transformers have various differences to CNNs which can give them advantages in certain tasks. Transformers are capable of capturing long-range pair dependencies between regions of an input histopathology patch or slide region. They can attend to all positions in the input image simultaneously, unlike CNNs which have limited receptive fields due to convolutional operations. This makes transformers ideal for tasks were understanding relationships between distant parts of an image are crucial. Transformers also have a crucial component called a self-attention mechanism. This enables them to weigh the importance of different parts of the input image when making predictions. This allows transformers to focus on relevant information.

Transformer architectures have been used for patch level supervised classification tasks similar to those mentioned in section 2.5.1. The benefit of using transformers is that they considers the relationship between all image regions, whereas CNNs are limited to the local context captured by their fixed-size convolutional filters. Ding et al. [48] combined a transformer with a CNN to extract the local features and global contextual information within tissue structures. They found superior performance over fully CNN architectures. Chen et al. [28] fine-tuned a pre-trained vision transformer on several histopathology classification tasks and found superior performance compared to a pre-trained ResNet-50 CNN model. Transformers can also be used for segmentation tasks similar to those mentioned in section 2.5.1. Nguyen et al. [106] evaluated several transformer segmentation models against CNN segmentation models for tumour segmentation. They found that transformers achieved better performance compared to the CNN counterparts.

One limitation of mentioned MIL approaches in section 2.5.2 is they do not consider the spatial relationship of patches within an image. Patches are often extracted and embedded into the feature space before being concatenated, thus eliminating any positional information. Transformers offer a solution by modelling the dependencies between patch embeddings using a self-attention module. This module computes a global attention matrix between each and every patch to determine how important one patch is to another. This works by computing a query, key and value vector for each patch in an image. The query represents the contribution of the current patch to the attention calculation, the key represents all other patch contributions to current attention calculation and the value represents the content information of the current patch. An attention score is calculated using the dot product between the current patch query and all other patch key vectors. The current patch feature is then updated by the weighted contribution of all patch values where the weights were determined by the attention mechanism. This is done for all patches in a slide or a selection of important patches to speed up computation. The self-attention mechanism therefore allows the features of a patch to be updated and influenced by all other patches in a slide. Additionally, positional encodings allow for learning of spatial relationships between patches which can allow for patches in closer proximity to influence the features of one another more than patches which are further away. Although transformers are powerful tools, they are vastly more computationally expensive compared to traditional MIL methods due to the greater number of model parameters and the exhaustive computations of attention values between all patches. Therefore, much of the research into transformers for WSI analysis focus on increasing the efficiency of such methods.

Huang et al. [66] proposed SeTranSurv which adopts a self-supervised learning approach based on Sim-CLR [32] for initial feature representation followed by a transformer encoder block to model dependencies between patch embeddings. They only sample 600 patches per image due to the large computational graph. Lu et al. [94] adopted a similar approach for Glioma sub-type classification which, instead of using random patches as inputs to a transformer, they jointly learned a instance-level classifier and the top-$R$ instances were used instead. Deformable Transformer MIL (DT-MIL) was introduced by Li et al. [82] which, unlike a traditional transformer, only attends patches to a small set of key instances instead of every other patch thus reducing memory and compute times. This is done by introducing a linear layer with learnable weights which determines a small set of positional offsets from a particular query. The query therefore only attends the the keys of the features are these offsets. TransMIL [122] uses the Nystrom method proposed in [159] to approximate the self-attention instead. Zhao et al. [175] proposed SETMIL which produces representations for sub regions within the WSI using a tokens-to-token transformer and then combines these sub regions globally. Chen et al. [30] introduced a Hierarchical Image Pyramid Transformer (HIPT) to capture morphological representations at different image scales. They additionally utilise the DINO framework [22] for self-supervised pre-training of the transformer blocks. DINO is similar to the other self-supervised techniques discussed above, however does not require negative examples which is particularly useful for histopathology slides where it is difficult to filter images to find true negative examples.

### 2.5.4 Transfer Learning

This is the technique of using a model which has already been trained using data from a different source domain and applying it to a task within a target domain. Generally, the source domain is related to the target domain so that any knowledge transferred to the target domain is applicable. The aims of transfer learning are to speed up training times as a model is already pre-trained from a different domain, increase performance by applying specialised features to a new task and finally to improve performance in cases where the sample size of the target domain is small. When there is plenty of training data for model development then training from randomly initialised weights instead of using a pre-trained model can be beneficial. This is because a model can learn task-specific features directly from the data without being biased by features learned from a different task.

Within Digital Pathology, transfer learning is typically done using ImageNet pretrained models such as VGGNet [128], ResNet [60] or InceptionNet [139]. Although ImageNet contains natural images many of the features learned from natural images can be applied to histopathology, especially from earlier convolutional layers which have been found to recognise edges, textures and shapes [177]. Often the pre-trained models are fine-tuned to a specific histopathology task by replacing the source classifier with a classifier trained on a specific task such as tumour prediction. However, the use of pre-trained models does not always guarantee better performance as discovered by Liu et al. [93] who found training from random initialisation produced better results than pre-training but at the cost of longer training times.

Transfer learning is adopted in this work for feature extraction due to availability of datasets and annotated data.

## 2.6 Explainable AI and their uses in Histopathology

With the increased use of deep learning methods such as CNNs and transformers for image analysis, concerns surrounding their interpretability and transparency have been raised. To address this a field of XAI methods has emerged with the goal of understanding model predictions and their failures. This is particularly important in areas of high-stakes decision making, such as healthcare. Here recent advances in the field of XAI for image analysis will be discussed and how such approaches have been adapted for use in computational pathology.

### 2.6.1 Taxonomy of Explainability

#### Post-hoc vs Ante-hoc

*Post-hoc* approaches focus on adding interpretable power to models after the model has been trained. The benefits of such an approach are that pretrained deep models can be investigated without needing to be retrained. *Ante-hoc* approaches focus on using or generating transparent models. These are models which are intrinsically interpretable by design. Simple models such as linear classifiers or shallow decision trees are examples of interpretable models.

#### Local vs Global

*Local* explanations provide insights into a single model decision for one particular instance. This could involve ranking the feature importance or relevance of pixels [118] for the classification of a single prediction. *Global* explanations on the other hand provide insights into the classification of a whole class. This helps identify the semantic features used for classifying a whole class [69] and detect biases in the data distribution.

#### Model Specific or Model Agnostic

Model *specific* methods provide explanations only for certain models. This often requires a specific model architecture or a component, such as a certain layer, for example a global average pooling layer [176]. Model *agnostic* methods provide explanations regardless of the architecture choice [114] allowing for explanations, regardless of architecture to be generated, making such methods more accessible.

### 2.6.2 Saliency Maps

Saliency maps were some of the first methods to attempt to dissect predictions from CNNs. Although the saliency methods discussed here are not used in this work, they are important to discuss as they provide a foundation for determining important visual information used by a model when making a classification.

The basic idea behind saliency maps is to assign a saliency value to each pixel or region in an image, indicating its relative importance. Saliency maps can be generated using different methods, but they generally involve analysing the gradients, activations, or responses of a DNN model that has been trained on large-scale image datasets. By examining the model's internal representations and understanding which input regions activate certain neurons or layers the most, saliency maps can provide insights into the visual cues that influenced a model decision.

*Gradient-Based* methods were introduced in 2013 by Simonyan et al. [127] who used the gradient of the output class score with respect to the input image pixels to determine which pixels were the most influential. An extension was introduced in Springenberg et al. [135] with Guided Backpropagation which prevented backward flow of negative gradients, corresponding to neurons in a network which decrease the activation of a specific class output. This was achieved by modifying how gradients are backpropagated through ReLU activation functions. Sundararajan et al. [138] further proposed integrated gradients to improve the visual appearance of the saliency maps. The method begins with a baseline image, such as a black image which produces a low output score for a particular class. Then the gradients of each pixel are computed along a straight line path between the baseline image and the input image. This essentially computes the area under a curve which models pixel gradient as a function of the pixel value. SmoothGrad was also proposed by Smilkov et al. [132] which is a smoothing of the vanilla gradient method by computing the average gradients across $N$ images with additional Gaussian noise.

*Class Activation Maps* (CAM) were first introduced by Zhou et al. [176]. The approach makes use of the feature maps following the final convolutional layer. The feature maps are weighted by the fully connected layer used for classification. Because of this, the feature maps must be pooled into a single value and therefore the approach is not model-agnostic. The final CAM for a particular class $c$ is given by

$$CAM_c(x) = \sum_k w_k^c f_k(x) \tag{2.4}$$

where $w_k^c$ are the weights for class c in the fully connect layer and $f_k(x)$ are the feature maps produced by an input image $x$. To make the approach model agnostic and additionally allow for understanding of CAMs from layers other than the final layer, Selvaraju et al. [118] introduced Grad-CAM. This approach uses the gradient information flowing into the convolutional layer to assign importance values to each neuron. For a feature map the gradients for each neuron are average pooled to get the importance score. One short-coming of Grad-CAM was its ability to localise multiple occurrences of an object in an image. To improve upon this Chattopadhay et al. [24] designed Grad-CAM++ which uses a weighted average of the gradients flowing through a feature map to determine the feature map importance score, as apposed to the average pooling of the original Grad-CAM.

CAMs have been used in digital pathology to visualise patch regions that correspond to mitosis. Then features such as cell count, colour and texture were extracted from the regions to understand why they were important for model predictions [137]. Mi et al. [100] deployed Grad-CAM for explaining patch predictions of an inception V3 model for breast cancer subtype prediction. Extension of Grad-CAM to WSI was explored by Brancati et al. [18] who first embedded patch features using a CNN. These patches where then arranged into a grid preserving their spatial arrangement. A 3D Convolutional Layer with two different attention modules produces attention feature maps which can be viewed using the Grad-CAM approach.

*Perturbation* methods involve altering different regions of an image and seeing how such perturbations impact the model output. The idea being that regions of the image which, when perturbed, cause the biggest change in the model output are the most important for the given prediction. Perturbation based approaches were introduced by Zeiler and Fergus [168] who occluded image regions using a grey square to understand how the model output changed. Local interpretable model-agnostic explanations (LIME) introduced by Ribeiro et al. [114] first segmented an image up into super-pixels which are regions of an input image which contain similar features. The image is then perturbed by replacing the values of each super-pixel with either a 0 or the average value of the super-pixel. The output of the model can then be measured for different arrangements of perturbed super-pixels. Randomized input sampling for explanation (RISE) was another perturbation technique proposed by Petsiuk et al. [110]. Such an approach produces a set of masks which are element-wise multiplied with the input image and the subsequent model predictions for each masked image are recorded. The saliency map can then be computed as a weighted sum of random masks, where weights are the probability scores that masks produce, adjusted for the distribution of the random masks.

LIME with multi-objective genetic algorithms was applied to lymph node metastases detection to generate explanations by evolving a segmentation to optimise three evaluation goals simultaneously [134]. Additionally, Graziani et al. [55] proposed Sharp-LIME which proposed superpixels using detected nuclei contours to make sure explanations aligned with medical concepts.

*Decomposition based* approaches identify parts of an image that directly provide evidence for a model decision unlike gradient and perturbation approaches which determine how much changing an image pixel alters the prediction. Layer-wise relevance propagation (LRP) [9] computes relevance scores for each neuron starting from the output neuron all the way to the input pixel space. The relevance score $R_i^l$ for a neuron $i$ in layer $l$ is determined by

$$R_i^l = \sum_j \frac{a_i w_{i,j}}{\sum_{i'} a_{i'} w_{i',j}} R_j^{l+1} \tag{2.5}$$

where $a_i$ is the activation of neuron $i$, $w_{i,k}$ is the weight connection between node $i$ and a neuron $j$ in layer $l+1$, $a_{i'}$ are the activations for all other neuron in layer $l$ and $w_{i',j}$ are the weight connections between all other nodes in layer $l$ and each neuron $j$ in layer $l+1$, finally $R_j^{l+1}$ is the relevance score of each neuron $j$ in layer $l+1$. The relevance scores for each pixel can then be used to generate the saliency map.

LRP was used for prediction of three different predictions tasks: cancer patch classification, lymphocyte patch prediction and whole-slide protein/gene expression prediction and was shown to be an effective strategy for cell localisation [16].

While saliency maps provide a visual explanation for what input pixels were important for a prediction these explanations can still be difficult to interpret. Often saliency maps are unclear and visually unappealing making them hard to understand. Additionally, once an image region is highlighted it is still unknown what this region might represent. If the face of a dog is highlighted in an animal classification task, was it the texture of the fur or the facial features which caused the final model prediction? To overcome this, concept attribution methods were introduced to provide more human-interpretable features.

### 2.6.3 Attention Mechanisms

Attention mechanisms help guide a network towards a decision by prioritising image regions which should be paid the most attention to and are used in chapters 5 and 6 in this work. Visualising the attention weights across the input image helps identify the most salient image regions and shares many similarity with the saliency methods discussed above. The use of attention in conjunction with MIL is regularly adopted for weakly supervised approaches in digital pathology [67, 84, 98] as already discussed in section 2.5.2. Attention is commonly used to weight each instance in the bag on its contribution to the label. Visualising the instances with the greatest attention weights gives some interpretable power to the model. For a bag $H = \{h_1, ..., h_K\}$ containing $K$ instance embeddings, the attention weight for the $k$th instance is given by,

$$\alpha_k = Softmax[w^T tanh(V h_k^T)] \tag{2.6}$$

where $w \in \mathbb{R}^{L \times n}$ and $V \in \mathbb{R}^{L \times M}$ are learnable parameters, $n$ is the number if classes, $L$ is the hidden layer dimension and $M$ is the instance embedding dimension. Other uses of attention attempt to discern a mask to guide the network towards the salient image regions. Instead of weighting the patches within a WSI this method highlights the salient regions within the pixel space [161, 65]. Other approaches use attention weights across the embedding space of patches from a WSI, using only WSI labels for training, similar to MIL [141, 18]. Some alternate uses for attention have also been proposed which produce a diagnostic report alongside the classification [173, 34]. This provides insight into a model's reasoning in an interpretable format for pathologists. For these approaches it is required that pathologists provide descriptive texts of sample input images in order to train a natural language model. Often these texts are required to contain specific information such as: state of nuclear pleomorphism, cell crowding, cell polarity, mitosis and prominence of nucleoli. This is done to keep diagnostic reports consistent and clinically relevant. LSTM modules [62] are used to generate the descriptions from visual features extracted by a CNN.

### 2.6.4 Concept Attribution

This style of explanation identifies human interpretable concepts which the model has learned and are predictive of certain model outputs. TCAV (Testing with Concept Activation Vectors) [69] introduced by Kim et al. is a model-agnostic explanation approach which identifies concepts within the latent space. These concepts are represented as vectors within the latent space. Image features which align along the direction of the concept vector are said to contain this concept. To learn these concept vectors a set of images which have been identified to contain this concept and a set of random images are input through a trained neural network. The model activations at a particular layer of interest are collected and a linear classifier is trained to determine if the concept image activations and random image activations can be separated. If the classifier can separate the classes then the model has identified a certain concept, and the coefficients of the linear classifier represent the concept vector.

For histopathology tasks, TCAV was extended to a regression based approach by Graziani et al. [57] for identification of learned concepts from pathology image patches. Values of concepts such as total nuclei area or nuclei texture are measured for images and a regression model is fit to predict these concepts using the activations from a particular layer in a model as input. The class sensitivity to each concept can be measured to determine the influence of the concept towards a prediction. The identification of concepts used for survival analysis by a vision transformer method were found by Shen et al. [124]. Important patches for a particular prediction were identified and nuclei were segmented from the patches. The nuclei were divided into two categories according to the patients' survival time lengths. For each group, features relevant for pathological interpretation are extracted and the probability distribution for every feature is measured. Finally, they devised a separability score between the distributions of each group to determine feature importance. These approaches determine pathologically relevant concepts used by neural networks to made decisions however, the explanations are still only an estimate of the decision making process. Instead, many approaches use the identified image concepts as input to a classification model, ensuring that the prediction is based solely upon the provided concepts. Although many earlier works made use of pixel and object level features for image classification for pathology [73] these concept features are difficult to interpret for humans and therefore do not make for convenient and practical explanations. Diao et al. [45] computed human-interpretable features (HIFs) to predict molecular phenotypes. Various cell and tissue types were identified and segmented across a range of WSI and features relating to the size, shape, area and spatial occurrence of both cells and tissue were discovered. Spearman correlations between HIFs and molecular phenotypes were calculated to determine relevant HIFs. For lung cancer survival outcome prediction, Wang et al. [147] extracted 22 shape and boundary features from identified tumour regions. They discovered that 15 of them were significantly associated with patient survival outcomes. Novel digital scores, such as the Abundance of Tumour Infiltrating Lymphocytes have also been discovered for Oral Squamous Cell Carcinoma diagnosis. The work by Shaban et al. [121] segmented both tumour regions and lymphocytes before quantifying the co-localisation of these components. In their more recent work, they applied a similar approach to head and neck squamous cell carcinoma [120] but this time looked at the tumour-associated stroma infiltrating lymphocytes score (TASIL-score). Many of these discovered features are more understandable compared to DNN features because they represent the underlying biology.

### 2.6.5 Prototypes and case-base reasoning

Prototype explanation methods attempt to explain a model decision by providing similar images from the dataset which produce a similar model response and are explanation methods used in chapters 4 and 6of this work. This is an intuitive explanation technique which is easily understandable by humans. These works are closely aligned with the classical form of case-based reasoning [71] but make use of the latent space of deep neural networks to determine similarities between inputs and prototypes. Li et al. [86] proposed a prototype network based on an autoencoder and prototype classification network. The autoencoder embeds the image into a vector representation and the prototype classification network measures the $L^2$ distance between this and a set of learned prototype vectors. A similarity score between the image and these prototypes is calculated using the $L^2$ distances where a smaller distance would result in a larger similarity score. These similarity scores are then input to a linear layer for model prediction. This approach was improved upon by ProtoPNet [26] which learns prototypes for objects within images rather than a prototype that summarises a whole image. This approach was able to produce more fine grained classifications. Nauta et al. [104] incorporated a decision tree with the ProtoPNet where each node in the tree was represented by a prototype vector and the image similarity to this was used to decide which direction to go down the tree. Rymarczyk et al. [116] further noticed that classes can share semantic similarities which are not modelled in the original implementation. Therefore, they introduced a data-dependent merge-pruning phase where prototypes from separate classes are merged together if they both are found to be similar to the other's class. Kim et al. [68] have additionally argued that prototypes alone are not enough and parts of the feature space where prototypical examples do not provide good explanations should also be found.

Prototype based methods have been used in WSI analysis, mainly to improve upon the common MIL approaches. Rymarczyk et al. [117] used the similarity scores between each prototype in ProtoPNet and a patch as the features for that patch. Most other methods just used the features from a pretrained model. By using ProtoPNet the patch features were more interpretable. However, the final model decision is based on the attention aggregation of these similarity scores instead of using the original reasoning of *this looks like that*. Other methods have looked at using prototypes for vocabulary-based MIL methods [165, 145] which function very differently to the case-based reasoning method of ProtoPNet. Therefore, work that explores the application of the original implementation of ProtoPNet to histopathology is still required.

## 2.7 Performance Metrics

Performance metrics provide objective measures to assess how well a model is performing and can help in comparing different models or tuning model parameters. The choice of performance metrics is dependant on the data and the task and is therefore an important consideration that must be made. The performance metrics used throughout this work will now be introduced, the rationale for use and the way in which the metrics are calculated will also be explained.

To calculate each metric the True positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) first need to be counted. TP refers to correct predictions of a positive class and TN refers to the correct predictions of the negative class. In the case of a tumour vs healthy tissue classifier, tumourous regions would be the positive class and healthy tissue would be considered the negative class. FP and FN refer to the incorrect prediction of the positive or negative classes respectively. In other words, if a tumour image is classified as healthy by a model, it would be considered a FN. A FP on the other hand would be a healthy image classified as tumour by a model. After counting the numbers for each of these, many performance metrics can be calculated.

**Accuracy (ACC)**is the most basic and intuitive metric, measuring the proportion of correct predictions over the total number of predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.7}$$

**Precision** is the proportion of TP over the total number of predicted positives. It focuses on the accuracy of positive predictions and is useful when the cost of false positives is high.

$$Precision = \frac{TP}{TP + FP} \tag{2.8}$$

**Recall** is the proportion of TP over the number of actual positives in the dataset. Recall measures a model's ability to detect for the positive class and is useful when the cost of false negatives is high.

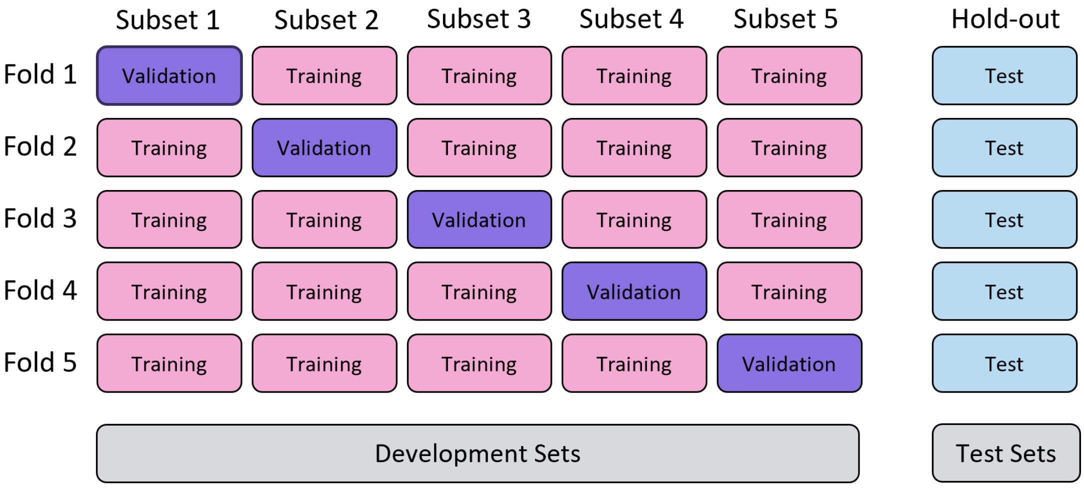
$$Recall = \frac{TP}{TP + FN} \tag{2.9}$$

**F1-score** is the harmonic mean of precision and recall. It provides a single metric to balance precision and recall. The F1 score is useful when you want to find an optimal balance between precision and recall. As false positives provide added stress on a healthcare system and false negatives cause patients to miss critical treatments, this metric is particularly important for medical diagnostic tasks.

$$F1score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \tag{2.10}$$

**Area Under the ROC Curve (AUC)** Plots the true positive rate (recall) against the false positive rate, providing an overall measure of performance across different classification thresholds.

Additionally, all models are trained and validated using K-fold stratified cross validation (Fig 2.2). This is a technique for estimating the generalisation ability of a particular model. In k-fold cross validation the dataset is split into k subsets. Then k models are trained where k-1 of these subsets are used for training and the remaining subset is used for validation. In the stratified version the proportion of each class is kept consistent between folds which is especially useful when working with unbalanced datasets. The main advantage of k-fold cross-validation is that it provides a more reliable estimate of a model's performance compared to a single train-test split. It helps in assessing how well the model generalises to unseen data and provides insights into stability and robustness.



|  | Subset 1 | Subset 2 | Subset 3 | Subset 4 | Subset 5 | Hold-out |
|---|---|---|---|---|---|---|
| Fold 1 | Validation | Training | Training | Training | Training | Test |
| Fold 2 | Training | Validation | Training | Training | Training | Test |
| Fold 3 | Training | Training | Validation | Training | Training | Test |
| Fold 4 | Training | Training | Training | Validation | Training | Test |
| Fold 5 | Training | Training | Training | Training | Validation | Test |
|  | Development Sets |  |  |  |  | Test Sets |

Figure 2.2: K-fold cross validation using 5 folds. The development set is split into 5 equally sized subsets. In each fold, one subset is used for validation and the rest are used for training. A different subset is used for validation in each fold. The whole out test set is identical across all folds.

# Chapter 3

# Whole-Slide-Images Datasets

## 3.1 Histopathology Slides

To produce histopathology slides, during a biopsy, tissue samples of interest are removed from the body. Tissue fixation is then performed to prevent autolysis and putrefaction. Autolysis is the complete destruction of tissue structures due to uncontrolled water and electrolytes dynamics in and out of the cell, and alteration in enzyme activity, while Putrefaction is the decay of organic matter by the action of microorganisms. Tissues are plunged into a fixation solution to prevent this from occurring. Next, samples are trimmed in order to reveal relevant surfaces and also to fit the cassette size. The tissue is then embedded, often using paraffin wax, to create a more supported structure to allow for thin slicing. For embedding, the tissue is first dehydrated and then cleaned to allow infiltration of paraffin wax. At the embedding centre the mould is filled with liquid paraffin wax and the tissue sample is removed from the cassette and is placed into the mould. Finally the mould is placed onto a cold plate to solidify the wax. When solidified, the tissue can be thinly sliced. Once sliced, the tissue sections are transferred to a warm water bath where they float to the surface and are then placed onto a slide under the water level. The warm water allows any wrinkles or creases in the tissue section to be smoothed out. Once the tissue is on the slide the supporting wax can be removed. The section is then stained to create contrast amongst the tissue structures, providing greater visibility

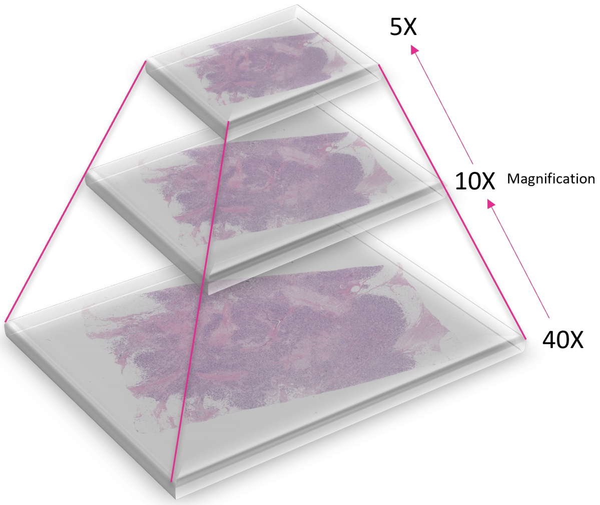


Figure 3.1: Pyramid structure of a digital pathology slide.

and ease of use. H&E is the routine staining and stains nuclei purple and proteins in pink [131]. Traditionally H&E stained images, or slide, are viewed under a microscope by trained pathologists for analysis and assessment of the tissue sample. However, digital pathology allows slides to be viewed on a computer screen. By doing this, slides are easier to view by allowing pathologists to zoom, pan, measure region sizes, annotate slides and share with one another. To view a slide digitally it must first be scanned by an appropriate scanner. This produces high-resolution images which can contain more than a billion pixels. Of course the whole image at this resolution cannot be viewed. Therefore, slides are often scanned at different magnifications and a pyramid structure is formed as shown in Fig 3.1. This allows the slide to be viewed at the appropriate magnification for the current zoom.

## 3.2 Dataset Descriptions

Throughout this work four histopathological datasets are used. Three of these datasets are open-source while the fourth was obtained through the Leeds Teaching Hospital Trust for which ethics approval was granted through Leeds WEST LREC reference no 05/Q1205/220. A brief description of each dataset will now be provided.

- **BACH**: Is the ICIAR 2018 Grand Challenge dataset for **B**re**A**st **C**ancer **H**istology Images [6] and is publicly available under the CC BY-NC-ND license. The dataset contains a total of 400 microscopy images with 100 images per class (4 classes). The four classes are Normal, Benign, In-situ carcinoma and Invasive carcinoma. The annotation was performed by two medical experts and images where there was disagreement were discarded. Each image is 2048 × 1536 pixels in size and the scale of each pixel is $0.42\mu$ m × $0.42\mu$ m. All images were acquired in 2014, 2015 and 2017 using a Leica DM 2000 LED microscope and a Leica ICC50 HD camera and all patients are from the Porto and Castelo Branco regions (Portugal).

- **Grade**: Is obtained from [47] and consists of 300 images from 21 patients with invasive ductal carcinoma of the breast of which 107 are grade 1, 102 are grade 2 and 91 are grade 3. The images are 1290 × 960 pixels in size. The image frames are from regions afflicted by tumour growth and were captured using a Nikon digital camera attached to a compound microscope with 40× magnification objective lens.

- **LTHT**: Is data obtained from Leeds Teaching Hospital Trust. The dataset contains 452 slides of primary tumours and lymph nodes from breast cancer patients. After processing the slides, 326 slides were available for model training and evaluation with 163 individual patients. The slides were scanned using an Aperio scanner at 0.25 micrometers-per-pixel using a 40× objective lens and are also FFPE. Metadata accompanies the slides and provides information on each patient's age, invasive tumour size, type, grade, stage, NPI, ER, PR and HER2 status to varying degrees.

- **TCGA-NSCLC**: Is data from The Cancer Genome Atlas (TCGA) open-access portal which contains a total of 1053 pathology slides and is available at https://portal.gdc.cancer.gov/. 541 of the slides contain lung adenocarcinoma (LUAD) and 512 contain lung squamous cell carcinoma (LUSC). Both of these subtypes are types of non-small cell lung cancer (NSCLC). Adenocarcinoma is the most common NSCLC, followed by squamous cell carcinoma. LUSC occur in the central part of the lung or the main airways (left or right bronchus) and develops in the flat cells that cover the surface of these airways. LUAD starts in the mucosal glands in the lining of these airways [167]. In this dataset There are a total of 956 patients, 478 in with LUAD and 478 with LUSC. All the slides were scanned using an Aperio scanner 0.25 micrometers-per-pixel using a 40× objective lens. The slides are Formalin-Fixed Paraffin-Embedded (FFPE) and not the Flash Frozen samples. Although this dataset does not contain breast cancer pathology slides, the dataset is very commonly used to compare performance in other works. Additionally, using this open source dataset allows for model training on a larger number of pathology slides, instead of only using the private LTHT dataset which has fewer slides.

## 3.3 LTHT Dataset Curation

For collection of the LTHT breast cancer dataset there are a range of tissue blocks available from the primary breast site and lymph nodes. One block was then selected from the primary site and one block of the lymph nodes was selected. This was done because it would not have been possible to process all available blocks for a patient. The selected primary tumour block was the most representative of the tumour as outlined by the original reporting pathologist. For the lymph node block, this was selected based on it being the most representative of the largest involved lymph node. In general, representative blocks contain the largest area of tumour across all available blocks.

After selection of the representative blocks the LTHT dataset contained a total of 451 slides from 174 patients. For each patient there were a range of slides of the primary breast site and lymph nodes stained with H&E, and in some cases IHC stained slides were available. As only the H&E stained slides were of interest, and not every patient had IHC stained images, the IHC slides were removed from the dataset. This resulted in 47 slides being removed. Following this, slides with missing metadata were also removed, this resulted in 3 slides being removed. Missing metadata refers to a slide not having any associated information which could be used to assign a label to the image for a particular classification task. In the case of the 3 removed slides they missed grade, NPI and Invasive tumour size information which were crucial for downstream analysis tasks. Next, patients who did not have both a primary site and lymph node image were also excluded. This resulted in the removal of 13 slides and 9 individual patients. Some of the remaining patients had more than one primary site or lymph node slide.

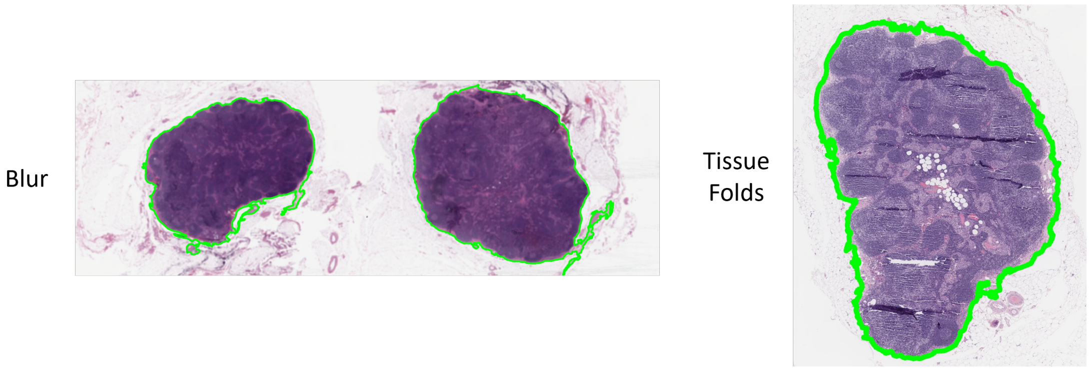


Figure 3.2: Slide Images with blur and tissue fold artefacts

As multiple slides were not available for all patients, one primary site and lymph node needed to be selected per patient. In all cases the additional slides were from the same lymph nodes or primary site tissue, but from a different slice. To remove duplicate slides first a quality control step was applied to remove slides containing excessive amounts of artefacts, such a pen marks or tissue folds. This only resulted in the removal of 2 slides, one for blur and another for a high quantity of tissue folds which can be seen in Fig 3.2. The slides belonging to one patient were also removed as some tiles within the primary site slide were corrupted; this removed 4 slides from the dataset. Random selection was then employed to select between the remaining duplicate slides per patient. Random selection removed a total of 58 slides. The final, full dataset description can be seen in Table 5.1.

## 3.4 WSI-Preprocessing

Pre-processing of the LTHT and TCGA-NSCLC datasets includes steps for tissue segmentation, patch creation and patch feature extraction. The steps for processing the BACH and Grade datasets will not be covered here but can instead be found in section 4.4.1 as these datasets are only used in that specific chapter, and because the datasets are open-source, most of the processing has already been done.

The aim of the WSI pre-processing is to remove redundant information from the tissue slide. Redundant information is any image region which provides little to no diagnostic value, and is not used by the pathologist when analysing a slide. The most obvious is slide background

Table 3.1: Patient Characteristics of all 163 patients included in the cohort.

| Patient Characteristics | Dataset |
|---|---|
| $n$ | 163 |
| **Invasive Tumour Type** ($n$ (%)) | |
| No Special Type | 115 (71) |
| Pure | 23 (14) |
| Mixed | 15 (9) |
| Lobular | 3 (2) |
| Other | 3 (2) |
| Unknown | 4 (2) |
| **Tumour Grade** ($n$ (%)) | |
| Grade 1 | 11 (7) |
| Grade 2 | 100 (61) |
| Grade 3 | 52 (32) |
| **Lymph Node Stage** ($n$ (%)) | 0.812 |
| Stage 1 | 0 (0) |
| Stage 2 | 137 (84) |
| Stage 3 | 25 (16) |
| **Tumour Size** $mm$ | |
| Median (Interquartile range) | 23.5 (16.0, 32.0) |
| **Nottingham Prognostic Index** | |
| Median (Interquartile range) | 4.8 (4.36, 5.50) |
| **Prognositc Group** ($n$ (%)) | |
| Good | 4 (2) |
| Moderate-1 | 44 (27) |
| Moderate-2 | 56 (34) |
| Poor | 46 (28) |
| Very Poor | 13 (8) |
| **Her2 Status** ($n$ (%)) | |
| Unknown<br>Negative<br>Borderline<br>Positive | 17 (10)<br>125 (77)<br>4 (2.5)<br>16 (10) |
| **ER Status** ($n$ (%)) | |
| Unknown<br>Negative<br>Positive | 14 (9)<br>17 (10)<br>132 (81) |
| **PR Status** ($n$ (%)) | |
| Unknown<br>Negative<br>Positive | 22 (13)<br>41 (25)<br>100 (61) |

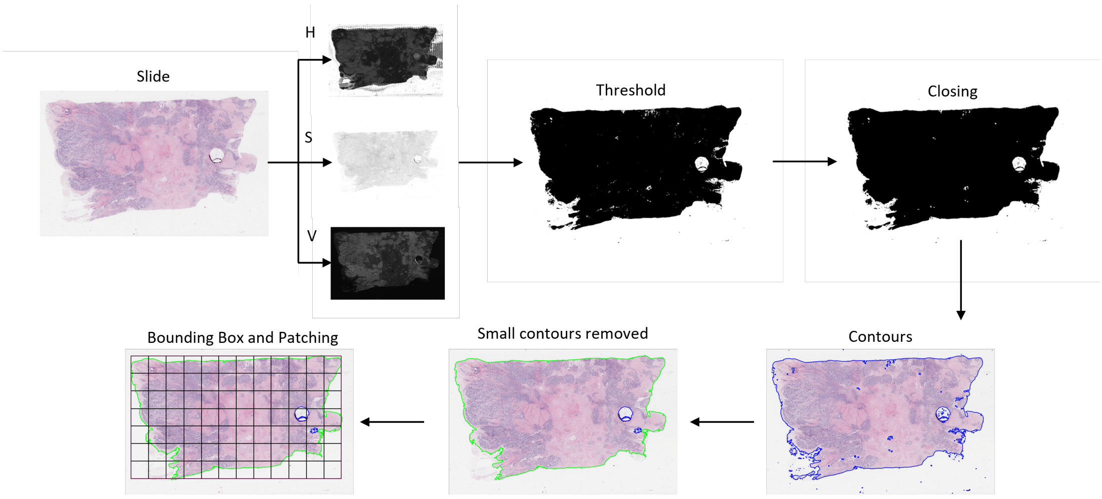


Figure 3.3: Tissue segmentation pipeline for WSI.

regions where there is no tissue present. Removal of these regions greatly reduces the amount of patches extracted from a slide and helps reduce memory requirements when training. Other less useful slide information could be large regions of fat tissue or much of the white space inside larger glandular structures. Fig 3.3 shows the pipeline for tissue segmentation and slide patching used for the LTHT dataset. Much of the processing was done following the steps from [98]. The processing steps are:

1. First a slide is downsampled by a factor of 64 and then converted from the RGB (red, green and blue) space to the HSV (Hue, Saturation and Value) space.

2. Median Blurring is applied to the saturation channel (S). This is a non-linear filtering technique which takes a median of all the pixels under the kernel area and replaces the central element with this median value. This is effective at reducing a certain type of noise like salt-and-pepper noise.

3. A binary threshold is then applied to the new S channel image where pixels above a thresholded value are set to 1 and set to 0 below this value. The optimal threshold was determined for each slide through visual assessment.

4. Morphological closing is applied to close small holes inside the foreground objects.

5. Contours are then found from the binary segmentation map. Contours are organised into

a two-level hierarchy. At the top level, there are external boundaries of the components (green in fig 3.3). At the second level, there are boundaries of the holes (blue in fig 3.3).

6. Contours of holes are removed if they are below a specified area.
7. For patch extraction, first a bounding box is fitted to the foreground contour.
8. The bounding box is then divided into non-overlapping patches and any patches where at least one patch corner is within the contour and the patch is not in a hole in the contour are kept. To test if a patch is in a hole the CV2 point polygon test is used. This function finds the shortest distance between a point in the patch and a contour. It returns the distance which is negative when the point is outside the contour, positive when point is inside and zero if point is on the contour. The point tested is the centre of the patch.

For processing the TCGA-NSCLC slides, a few additional processing steps were added to the original pipeline. This is due to the presence of pen marks found on these slides. Although some pen marks are present in the LTHT dataset, they are extremely rare and often small and do not obscure important information. However, this is not the case in the TCGA-NSCLC dataset where there are many slides with large tissue regions obscured. In order to remove pen marks different channels needed to be thresholded depending on the pen colour. This meant that each slide needed to be treated individually to avoid collecting patches with pen marks. For example, red pen marks were possible to remove by thresholding along the Green colour channel, while blue pen marks could be removed using the hue channel as shown in Fig 3.4.

Patches are extracted from the segmented tissue region in a non-overlapping manner meaning that no tissue regions appear more than once in the dataset. Patches are extracted at a size of 1024 pixels by 1024 pixels at the highest magnification of 40×. This is equivalent to 512 pixels at 20× and 256 and 10×. To extract features from the selected patches, they are fed in batches to a ResNet-50 [60] pretrained on ImageNet. The network parameters are frozen and the final classification head is removed. This results in a 1024 length vector being produced for each patch following the global average pooling over all pixels of the final feature maps. The features vectors are then concatenated together. Therefore, if a slide has $n$ patches, the final representation of the slide is an $n \times 1024$ array.

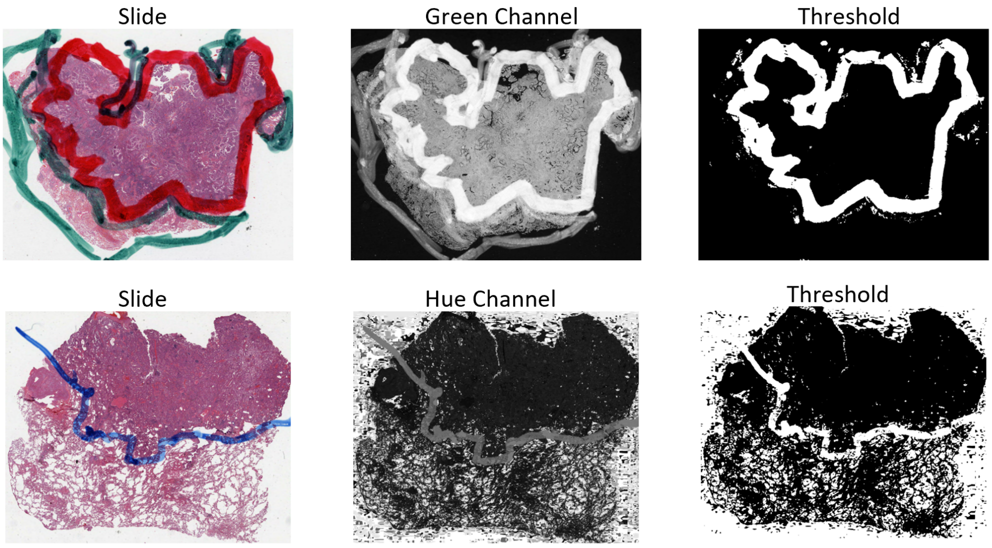


Figure 3.4: Channel thresholding approaches to removing pen marks.

# Chapter 4

# Validating Prototypical Part Networks for Interpretable Diagnosis

## 4.1 Introduction

In this chapter an interpretable methodology for breast cancer histopathology image analysis is investigated. Specifically a prototype-based neural network is utilised for classification of breast cancer sub-type and grade using regions of interest (ROI) from histopathology slides. The network used is a prototypical-part network (ProtoPNet) [26] and was designed as an intrinsically interpretable fine-grained image classification model. Fine-grained classifiers aim to distinguish between image classes that belong to the same overarching category but have subtle differences. The network incorporates novel prototype layers which allow for easier inspection of a classification. The methodology was inspired by the way humans explain things by comparing them and measuring their similarities. The model is therefore able to identify several parts of an input image *this* part of the image looks like *that* prototypical part of a particular class. Variations of ProtoPNet have been used for various medical imaging tasks, however not in the domain of digital pathology. Here, the prototype-based framework is assessed in the context of breast cancer digital pathology. In addition, two adjustments are made to the original approach to increase performance. First, an additional term is added to the loss function in the form of an orthogonality loss to encourage greater diversity amongst prototypes. This additional loss term was chosen as it has previously been shown [146] to create a more disentangled embedding space for natural image classification tasks. Due to the heterogeneity of histopathology images it is crucial that the prototypes can capture a variety of morphological structures. Second, in the original implementation Max pooling is used to determine the final similarity between a prototype and the input image however here this is replaced with an average pooling operation. Structures within a histopathology image can vary in shape and size between the images and may occupy different spatial locations altogether. It is therefore crucial to determine the average of the similarity scores rather than the maximum similarity to one spatial location.

## 4.2 Related work

Prototype-based neural network approaches have already been used in other medical image analysis tasks and have seen success in increasing model interpretability while not sacrificing any performance compared to their black-box counterparts. In X-ray image analysis a Negative-positive prototypical part network (NP-ProtoPNet) variant was used with a dataset of frontal chest X-ray images of Covid-19 patients, pneumonia patients and normal persons [130]. The model can not only detect positive evidence for a class but also rejects a class whose images do not have any part that matches with any prototype. XprotoNet was also designed for X-ray analysis [70] which can additionally predict the area where disease is likely to appear and compares the features in the predicted area with the prototypes. This allows the model to provide local explanations of the form *this looks like that* but also global explanations by determining critical image regions across a whole class. In digital mammography a revised ProtoPNet was used for classification of Mass Lesions in Digital Mammography [10]. Here domain knowledge was injected into the pipeline through domain expert annotations which helped refine the prototype search regions. Specifically prototypes are penalised for producing an activation map that does not correspond with these annotated regions and therefore produces clinically relevant prototypes.

## 4.3 Methodology

The model architecture of ProtoPNet will now be introduced along with several adaptations made to the original approach. The training scheme is also discussed.

### 4.3.1 ProtoPNet architecture

The network architecture consists of a backbone CNN $f$ with additional layers to compress the output to the dimension of the prototype vectors, followed by a prototype layer $g_p$ and a fully connected layer $l$ with weight matrix $w_l$ and no bias. The additional layers consist of a $1 \times 1$ convolutional layer with output dimension $c$, followed by a ReLU activation layer, an identical convolutional layer and finally a sigmoid activation function. For an input image $x$ the output of the CNN are a set of features maps of shape $h \times w \times c$ where $h$ and $w$ are the height and width and $c$ is the dimensionality of the features at each pixel position. These feature maps are then input into a prototype layer which is shown in Fig 4.2. This layer learns $m$ prototypes $P = \{p\}_1^m$. Each class is represented equally across these $m$ prototypes. The shape of these prototypes is $h_p \times w_p \times c$ where the channel features is the same as the feature maps but $h_p$ and $w_p$ are both equal to 1. The significance of the smaller prototype vectors is that each prototype can represent some prototypical activation pattern in a patch of the convolutional output. Hence each prototype can be thought of as some latent representation of a prototypical part of an input image. From the feature maps *f(x)*, the prototype layer $g_p$ computes a similarity score between each prototype and every patch location in the feature map. This creates a similarity map for each prototype of size $h \times w \times 1$. Each similarity map can then be pooled (e.g max or average pooling) to reduce the map to a single similarity score for each prototype. A large similarity score means that there is a patch in the features maps that is very close to a particular prototype in the latent space, and thus there is a patch in the input image that has a similar concept to the concept in that prototype vector. After the similarity score for each prototype is calculated they are passed to a fully connected layer $l$ which produces the output logits which are then normalised using softmax function to return the predicted probabilities for a given image belonging to a certain class. The input size of the fully connected layer is equal to the number of prototypes. The full architecture can be seen in fig 4.1.

### 4.3.2 Similarity Scores

To compute the similarity between the prototype vectors $P$ and the features maps $z = f(x)$ several different distance metrics were adopted to determine the optimal measure. Mathematically, for a prototype vector $p_j$ the prototype layer computes,

$$g_{p_j}(z) = log\left(\frac{(d_j + 1)}{(d_j + \epsilon)}\right) \tag{4.1}$$

where $d_j$ is a measure of distance between the prototype vector and each spatial location in the feature map $z$. The goal of equation4.1 is to return a larger similarity score when the distance between a prototype vector and a spatial location is small. A small distance between the prototype vector and spatial location means that they represent similar concepts in the latent space. Therefore, when this is the case, comparing the two should result in a larger similarity score. Two similarity measures as explored here. The first is the squared $L^2$ distance which is given by,

$$d_j^{L^2} = \|z - p_j\|_2^2 \tag{4.2}$$

and the second is the cosine distance given by,

$$d_j^{cos} = 1 - \frac{z \cdot p_j}{\|z\|\|p_j\|}. \tag{4.3}$$

The cosine distance given in equation 4.3 is 1 minus the cosine similarity. This is to make sure the function has a similar behaviour to the $L^2$ distance, where, as the two vectors move apart, the distance measure increases. The output for each prototype from the prototype layer is then pooled to produce a single similarity score for each prototype. As larger similarity scores mean that a prototype vector is similar to some region within the input image, max pooling is commonly used. However, average pooling is also explored here to determine the optimal approach. Min pooling is not often used as regions of the image not similar to a prototype are generally not of interest.

### 4.3.3 Training Method

The method of training a ProtoPNet is divided into three parts: 1) stochastic gradient descent (SGD) of layers before the last layer; 2) pushing of prototypes for visualisation; 3) optimisation of last layer.

#### 1) SGD of layers before the last layer

Here a semantically meaningful latent space is discovered and involves the optimisation of both the backbone CNN and the prototype vectors, the final linear layer weights are frozen during this step. The prototype vectors are initialised as a vector of random values. The goal is to encourage training images to have some latent patch which is close to a prototype representing its own class and further encourage prototypes of different classes to be separated in the latent space. To achieve this a loss function with three terms is proposed. The first term is the cross entropy loss which penalises incorrect model predictions for $n$ input samples.

$$\mathbb{L}_{ce} = min_P \frac{1}{n} \sum_{i=1}^{n} CrsEnt(l \circ g_p \circ f(x_i), y_i) \tag{4.4}$$

The second is the cluster cost which encourages each training image to have some latent patch close to a prototype of its own class.

$$\mathbb{L}_{cls} = \sum_{i=1}^{n} min_{p_j \in P_{y_j}} min_{z \in patches(f(x_i))} \|z - p_j\|_2^2 \tag{4.5}$$

The third term is the separation cost which forces prototype vectors of different classes to occupy different regions of the latent space.

$$\mathbb{L}_{sep} = -\sum_{i=1}^{n} min_{p_j \notin P_{y_j}} min_{z \in patches(f(x_i))} \|\tilde{z} - p_j\|_2^2 \tag{4.6}$$

However, training this way produced prototypes that eventually collapsed into a single prototype which represented the entire class. This therefore, reduced the diversity of the prototypes. To resolve this, similar to other work [146], an orthogonality loss was included into the function

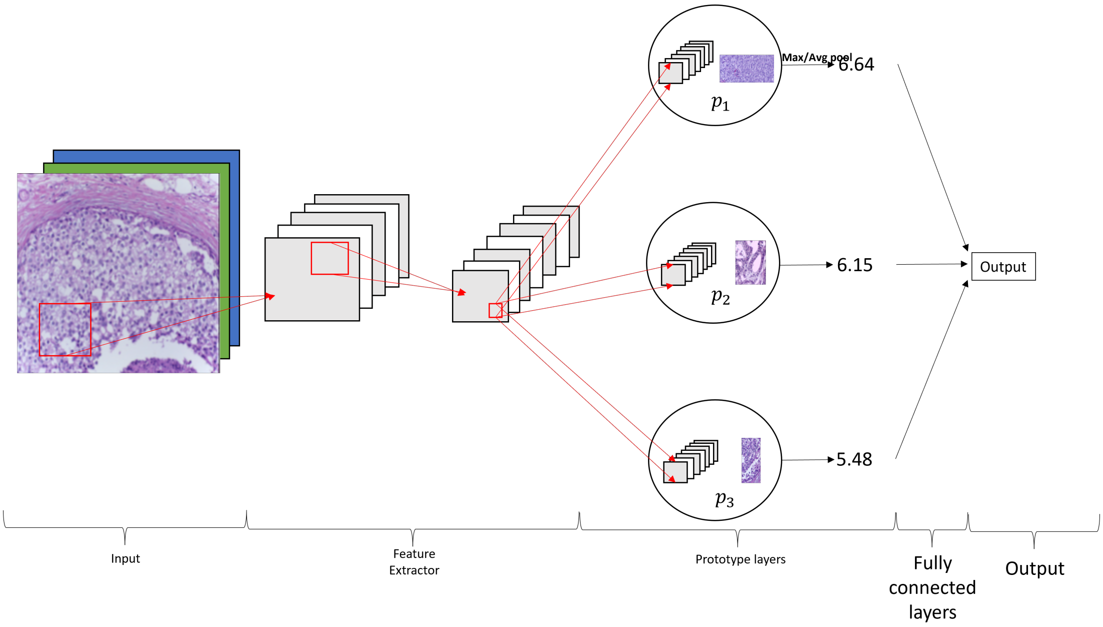


Figure 4.1: ProtoPNet architecture.

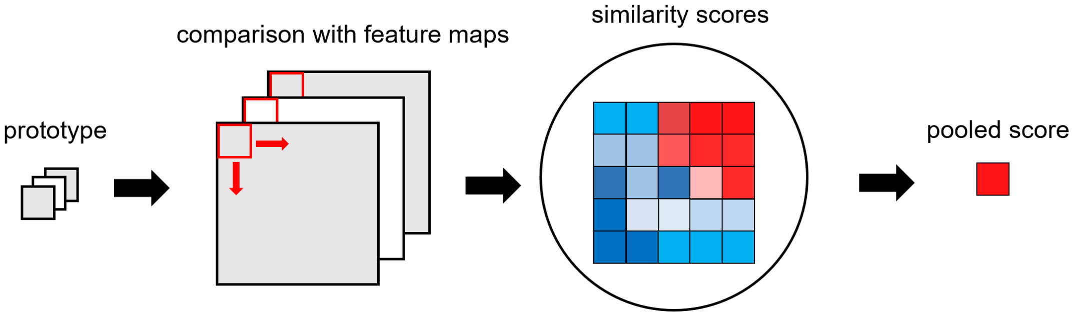


Figure 4.2: Prototype layer.

which creates greater diversity between prototypes. For classes $C$ this loss is given by,

$$\mathbb{L}_{orth} = \sum_{c=1}^{C} \|P_c P_c^\top - I\|_2^2 \tag{4.7}$$

where $I$ is the identity matrix of size $m_c \times m_c$ where $m_c$ is the number of prototypes assigned to each class. Therefore, the orthogonality loss encourages diversity within the prototypes for each class and helps prevent prototype collapse.

A weighted combination of these produces the final loss,

$$\mathbb{L} = \lambda_1 \mathbb{L}_{ce} + \lambda_2 \mathbb{L}_{cls} + \lambda_3 \mathbb{L}_{sep} + \lambda_4 \mathbb{L}_{orth} \tag{4.8}$$

### 2) Pushing of Prototypes for Visualisation

The goal of pushing prototypes is to equate each prototype with a training image patch so that they accurately represent image features found within the dataset. After several iterations of optimisation by SGD the prototypes should have taken on some semantic representation and can be visualised. To do this each prototype $p_j$ is *pushed* onto the nearest latent training patch from an image of the same class as assigned to the prototype. In practice this involves first scanning a prototype across the convolutional output of each input image from the same class as the prototype. This means moving the prototype through each possible position in the convolutional output and calculating the similarity score between the prototype and each position. For a convolutional output of $7 \times 7 \times c$, where $c$ is the number of features in the output feature map, 49 similarity scores would be produced, one for each of the 49 possible position in the $7 \times 7$ grid. Next the image which contains the latent patch with the highest similarity score, calculated using equation 4.1, is selected. The prototype vector is then set equal to the values of this latent patch. This ensures that the prototype is equal to a semantic feature within the training dataset and not just a similar to one. From the original image the similarity scores between the prototype and each latent patch is found again which creates a heatmap over the latent space. By upsampling this back to the input pixel space a heatmap over the input image can be visualised. The concept captured by $p_j$ can be visualised by the smallest rectangular patch of $x$ that encloses the pixels whose corresponding activation value in the upsampled activation map from $p_j$ is at least as large as the 95th-percentile of all activation values in that same map.

### 3) Optimisation of Last Layer

In the final connected layer $l$ the weight matrix $w_l$ contains entries $w_l^{(c,j)}$ corresponding to the $j$-th prototype of class $c$. For a specific class $c$ the value in the weight matrix is 1 for all $j$ with $p_j \in P_c$ and the value is -0.5 for all $j$ with $p_j \notin P_c$. Intuitively, prototypes belonging to a class $c$ should increase the predicted probability that the image belongs to that class. Conversely, the negative connection between a prototype not of class $c$ and the class $c$ logit means that similarity to a non-class $c$ prototype should decrease the predicted probability of class $c$. To optimise the last layer the backbone CNN parameters and prototype vectors are fixed and the loss function,

$$min_{w_l} \frac{1}{n} \sum_{i=1}^{n} CrsEnt(l \circ g_p \circ f(x_i), y_i) + \lambda \sum_{c}^{C} \sum_{p_j \notin P_c} \mid w_l^{(c,j)} \mid \tag{4.9}$$

is optimised. The second term is the $L1$ norm which prevents the final layer weights getting too large.

## 4.4 Experiments

The Prototypical Part Network is evaluated in the context of digital pathology on two different classification tasks within breast cancer. The first task is prediction of cancer type between Normal, Benign, in-situ carcinoma and Invasive carcinoma. The second task is breast cancer grading between low (grade 1), moderate (grade 2) and high grade (grade 3).

### 4.4.1 Datasets

For classification of breast cancer type the BACH dataset is used [6]. The dataset contains a total of 400 microscopy images with 100 images per class (4 classes), example images for each class can be seen in Fig 4.3. Each image is $2048 \times 1536$ pixels in size and the scale of each pixel is $0.42\mu m \times 0.42\mu m$.

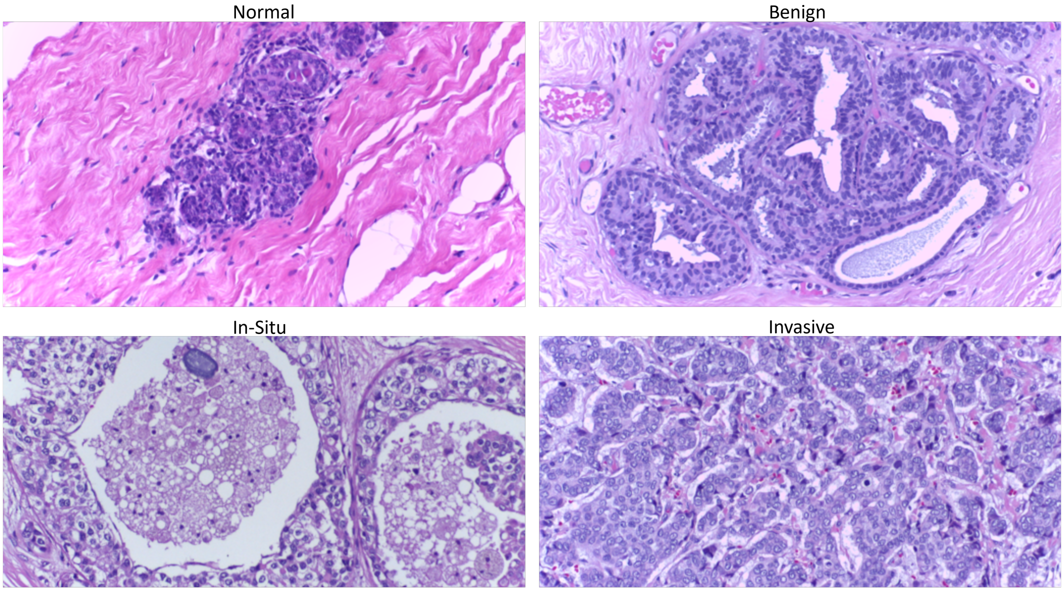


Figure 4.3: Examples of images from the BACH dataset. One image from each class is represented.

The breast cancer grading dataset was obtained from [47] and consists of 300 images of which 107 are grade 1, 102 are grade 2 and 91 are grade 3. The images are $1290 \times 960$ pixels in size and correspond to 21 different patients with invasive ductal carcinoma of the breast. The image frames are from regions afflicted by tumour growth and were captured using a Nikon digital camera attached to a compound microscope with 40× magnification objective lens. An example image from each class can be seen in Fig 4.4.

### 4.4.2 Data augmentation

Data augmentation was adopted to increase the size of the datasets and improve model training. To do this the Python package Augmentor [8] was utilised. For each image in a dataset the image was flipped along the horizontal, vertical or both axis. Then the flipped images were rotated by a maximum of 20 degrees in either direction. Additionally, the flipped images were also sheared by a maximum of 10 degrees. This created 11 different views per image. Therefore the BACH dataset was extended to a size of $4,400$ and the grading dataset to a size of $3,300$

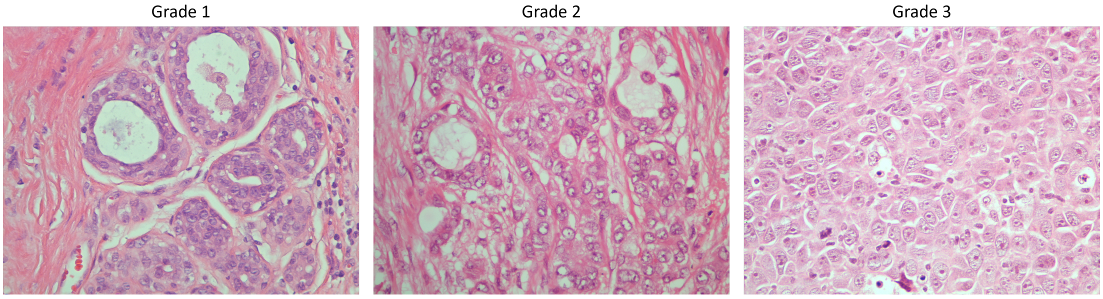


Figure 4.4: Examples of images from the grading dataset. One image from each class is represented.

images. During the prototype visualisation step (section 4.3.3) only the original images are used to prevent prototypes being discovered from similar images.

### 4.4.3 Implementation details

Several different backbone architectures were tested to determine the impact of architecture on prototype discovery and performance, these included: ResNet-18 [60] and VGG11[128]. Adam was used for optimisation of the network parameters with a weight decay of $1 \times 10^{-3}$. The learning rate for the backbone architecture was $1 \times 10^{-4}$ and $3 \times 10^{-3}$ for the prototype vectors. For the loss function coefficients $\lambda_1 = 1$, $\lambda_2 = 0.8$, $\lambda_3 = -0.08$. These parameters were chosen as they are the recommended parameters from [52]. A range of values $\{0.2, 0.4, 0.6, 0.8, 1.0\}$ for $\lambda_4$ were evaluated and the value with the highest validation AUC was validated on the test set. The value of $\epsilon$ in eq 4.1 was set to $1 \times 10^{-4}$. Prototypes were set to a size of $1 \times 1 \times 128$ and 10 were assigned to each class. Therefore, 40 prototypes were used for breast cancer sub-typing and 30 prototypes were used for breast cancer grading. All images were resized to a size of $256 \times 256$ pixels. The CNN backbone features and prototype vectors were trained for a maximum of 30 epochs. Prototypes were pushed every 10 epochs which is the default frequency recommended from the original implementation [26]. This frequency is chosen to strike a balance between updating the prototype vector values and making sure they still represent real concepts within the latent space. If the prototype vectors are pushed too often, the vector values are unlikely to have changed enough for the prototype to be pushed to a new latent training patch. This would greatly slow down training. Alternatively, if the pushing is too infrequent the prototype vector values could change too much. When they are eventually pushed, the prototype vectors could be very homogeneous. The final classification layer was trained after prototypes were pushed for 2 epochs. 5-fold cross validation was used and the model with the highest AUC on the validation set during the final layer optimisation phase was selected for test set evaluation. Before training a warm up phase is utilised which trains only the prototype vectors and the additional layers used for compressing the CNN features maps to the same number of features as the prototype vectors. The warm up phase lasts for 10 epochs and the learning rate is again $3 \times 10^{-3}$. Adam is used for optimisation during the warm up phase as well.

All prototypical part networks were compared against baseline models. These models used the same backbone CNN feature extractor (ResNet-18 and VGG11) however the prototype layers were excluded. Instead, the final features maps are passed through a global average pooling layer following by a linear classifier. For consistency, Adam was used for optimisation of the network parameters with a weight decay of $1 \times 10^{-3}$ and a learning rate of $1 \times 10^{-4}$ used. Cross Entropy loss was also used. The baseline models were also trained for a maximum of 30 epochs.

## 4.5 Results and Discussion

### 4.5.1 BACH

Model performance on the BACH dataset can be seen in Table 4.1 and Table 4.2 for the ResNet-18 and VGG11 backbones respectively. The results show that the use of the prototypical part network with average pooling and either distance metric outperforms both baseline methods. The ResNet-18 and VGG11 baselines achieved test AUCs of $0.935 \pm 0.039$ and $0.913 \pm 0.040$ respectively whereas the best performing ProtoPNet models achieved $0.948 \pm 0.005$ with a ResNet-18 backbone and $0.944 \pm 0.005$ with a VGG11 backbone. The ResNet performed best using the cosine similarity metric while the VGG11 performed best with the Euclidean distance similarity metric. However, the different similarity metrics did not produce significantly different results overall. The best performing model with the ResNet-18 backbone using Euclidean distance achieved a test AUC of $0.943 \pm 0.023$ compared to the $0.948 \pm 0.005$ test AUC of the

Table 4.1: Comparison between baselines and ProtoPNet with ResNet-18 backbone on the BACH dataset. Validation and Test AUC and accuracy (ACC) ± standard deviation are presented across 5-folds

| Model | pooling | sim metric | Val AUC | Test AUC | Val ACC | Test ACC |
|---|---|---|---|---|---|---|
| Baseline | N/A | N/A | 0.963±0.019 | 0.935±0.039 | 0.827±0.087 | 0.775±0.071 |
| ProtoPnet | avg | Euclidean | 0.969±0.012 | 0.943±0.023 | 0.873±0.015 | 0.825±0.029 |
| ProtoPnet | max | Euclidean | 0.856±0.044 | 0.842±0.037 | 0.694±0.085 | 0.670±0.069 |
| ProtoPnet | avg | cosine | **0.972±0.004** | **0.948±0.005** | **0.882±0.007** | **0.827±0.018** |
| ProtoPnet | max | cosine | 0.912±0.011 | 0.874±0.014 | 0.736±0.047 | 0.685±0.036 |

Table 4.2: Comparison between baselines and ProtoPNet with VGG11 backbone on the BACH dataset. Validation and Test AUC and accuracy (ACC) ± standard deviation are presented across 5-folds

| Model | pooling | sim metric | Val AUC | Test AUC | Val ACC | Test ACC |
|---|---|---|---|---|---|---|
| Baseline | N/A | N/A | 0.931±0.028 | 0.913±0.040 | 0.775±0.075 | 0.733±0.077 |
| ProtoPnet | avg | Euclidean | 0.967±0.013 | **0.944±0.005** | 0.857±0.027 | 0.802±0.013 |
| ProtoPnet | max | eudlidean | 0.888±0.019 | 0.850±0.019 | 0.714±0.035 | 0.625±0.032 |
| ProtoPnet | avg | cosine | **0.967±0.010** | 0.939±0.009 | **0.869±0.26** | **0.822±0.010** |
| ProtoPnet | max | cosine | 0.919±0.011 | 0.883±0.010 | 0.766±0.013 | 0.698±0.016 |

cosine distance counterpart. For the VGG11 backbone the cosine distance variant achieved a test AUC of $0.939 \pm 0.009$ compared to the $0.944 \pm 0.005$ test AUC achieved by the Euclidean distance variant.

The results do clearly show the benefit of the average pooling function over the max pooling function as there is an $\approx 9\%$ average increase in test AUC score between the maxpool and average pool counterparts. The results also show that the choice of backbone does not produce significantly different performance. The optimal values for $\lambda_4$ for training models on the BACH dataset can be seen in Table 4.3. There values were selected based on the highest average AUC score on the validation sets across the 5 folds.

Fig 4.5 shows the confusion matrices on the test set for a model trained on each fold. The models all had the ResNet-18 backbone with average pooling and cosine similarity metric. Across all folds normal and in-situ classes had the highest average recall of 0.898 and 0.889

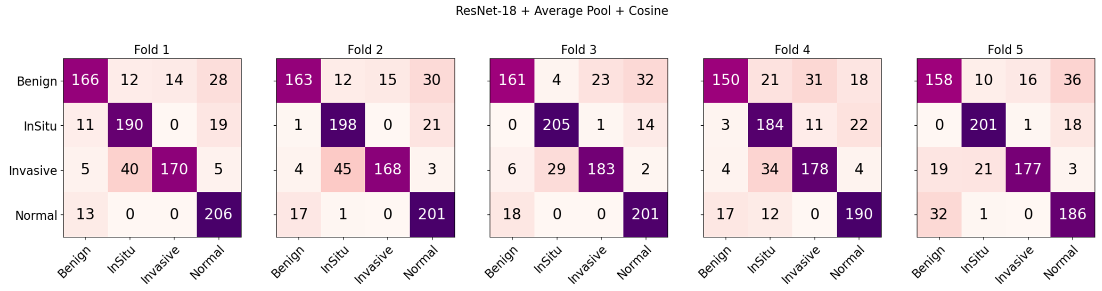


Figure 4.5: Confusion Matrices on the BACH test set for a ProtoPNet with ResNet-18 backbone, average pooling and cosine similarity metric trained from each fold.

Table 4.3: Orthogonality loss coefficients ($\lambda_4$) for BACH dataset

| config | ResNet-18 | VGG11 |
|---|---|---|
| avg + Euclidean | 0.6 | 0.8 |
| max + Euclidean | 0.4 | 0.8 |
| avg + cosine | 0.8 | 1.0 |
| max + cosine | 0.2 | 1.0 |

respectively. The benign class had the lowest with an average recall of 0.726. The invasive class had an average recall of 0.817. Although this is lower than the recall of both normal and in-situ classes, the majority of misclassifications are classified as in-situ. Meaning that, although the sub-type is incorrect, the model still detects malignancy. For the benign class the majority of misclassifications are for the normal class.

Invasive and benign classes had the highest average precision with 0.889 and 0.844 respectively whereas the normal class had the lowest average precision of 0.794.

As the model is intrinsically interpretable, the similarities between the input image and each prototype can be compared and ranked in order of similarity scores. This allows for visualisation of the decision making process. Fig 4.6 shows the similarity between the 3 most similar prototypes to the input image of invasive carcinoma. The class was correctly predicted by the model and all 3 top prototypes are from the invasive class. The prototypes all capture regions of tumour with high degrees of nuclear pleomorphism. The third prototype additionally captures small regions of fat which are also present in the input image. The prototype similarity

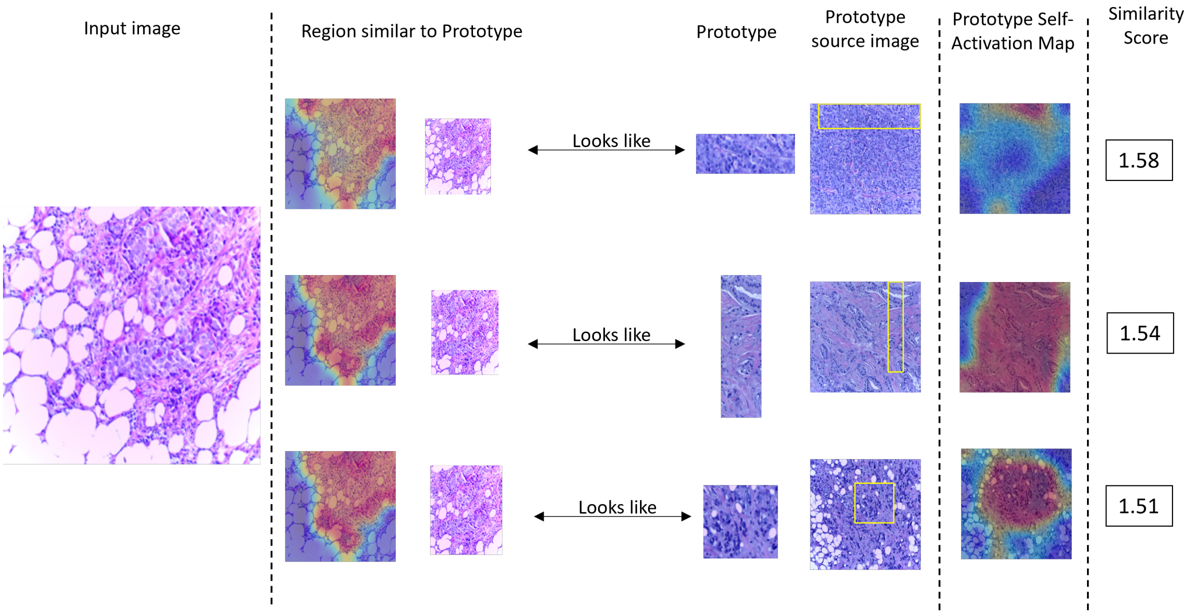


Figure 4.6: Explanation for model prediction of Invasive Carcinoma showing the three most similar prototypes to the test image. From left to right, the first image is the input image, the next is a heatmap showing the image region most similar to a prototype, the third is a crop from this heatmap, the fourth is a crop which represents the prototype, the first image is the training image the prototype was taken from, the final image is the activate map of the prototype across the source training image. The similarity score between each prototype and the input image is shown to the far right. The most similar prototype to the input image is at the top of the figure and the third most similar is at the bottom.

heatmaps all activate over the invasive tumour regions and do not activate over the large area of fat in the bottom left of the input image. Fig 4.7 shows the explanation for an in-situ case, again the model predicts the class correctly and the top 3 most similar prototypes contain in-situ carcinoma.

Fig 4.8 shows the 4 prototypes with the highest similarity score from each class for an input image from the Benign tissue class. The top row shows each prototype's activation map with the input image. The bottom row shows each prototype's similarity map from the source image they came from. For each column the model suggests that the red regions in each of the two images are semantically similar. For the benign class the prototype is sourced from a single duct and this is similar to all ducts within the lobule present in the input image. For the normal class the prototype with the highest similarity focuses on the interlobular stroma in both the

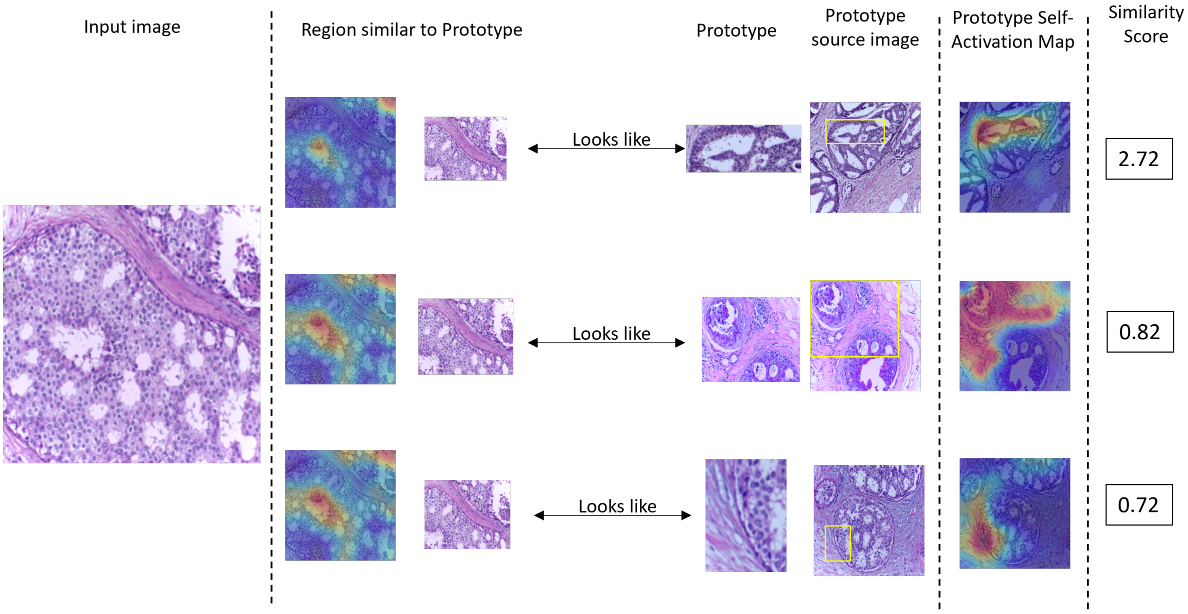


Figure 4.7: Explanation for model prediction of in-situ Carcinoma showing the three most similar prototypes to the test image. From left to right, the first image is the input image, the next is a heatmap showing the image region most similar to a prototype, the third is a crop from this heatmap, the fourth is a crop which represents the prototype, the first image is the training image the prototype was taken from, the final image is the activate map of the prototype across the source training image. The similarity score between each prototype and the input image is shown to the far right. The most similar prototype to the input image is at the top of the figure and the third most similar is at the bottom.

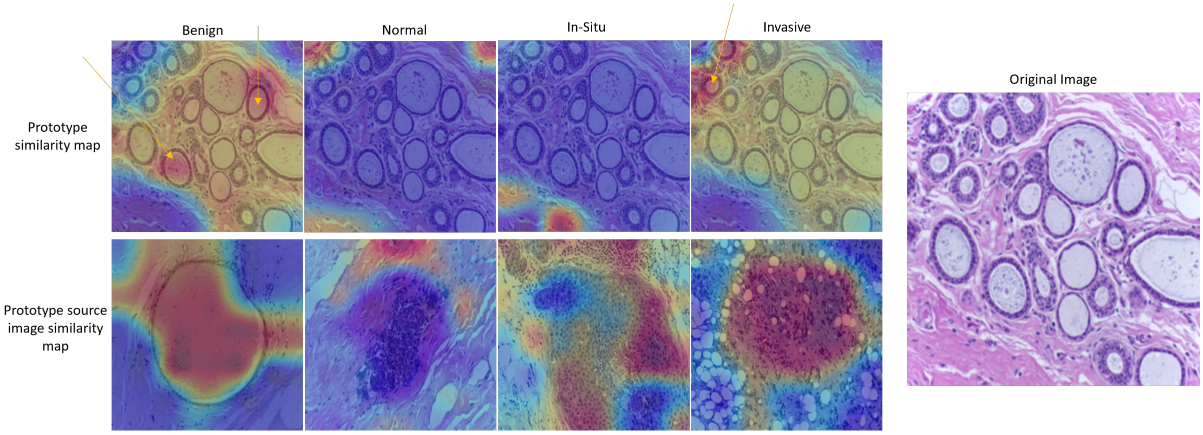


Figure 4.8: Examples of similarity heatmaps between the most similar prototype from each class and an image from the Benign Tissue class. Top row shows similarity maps between prototypes and the test tissue while the bottom shows shows the similarity maps between prototypes and their respective source image.

Table 4.4: Comparison to other methods on the BACH dataset

| Paper | Method | Test AUC | Test ACC |
|---|---|---|---|
| Araujo et al. (2017) [5] | CNN + SVM | - | 0.778 |
| Wang et al. (2018) [152] | VGG16 + SVM | - | 0.830 |
| Vo et al. (2019) [143] | Model Ensemble | - | 0.964 |
| Lu et al. (2019) [96] | Attention Based MIL + self-supervised feature learning | $0.968 \pm 0.022$ | $0.950 \pm 0.027$ |
| Li et al. (2019) [89] | ResNet-50 + SVM | - | 0.950 |
| Yang et al. (2020) [161] | Guided Soft-Attention | - | 0.930 |
| Yan et al. (2020) [162] | Inception-V3 + Bidirectional LSTM | - | 0.913 |
| Zhang et al. (2021) [171] | Multi resolution attention sampling | - | 0.850 |
| Li et al. (2021) [81] | Multi-view Attention-Guided MIL | 0.990 | 0.936 |
| Ours | ProtoPNet | $0.948 \pm 0.005$ | $0.827 \pm 0.018$ |

input image and the source image. This suggests that there is likely little evidence in the input image to support the normal classification. For the invasive class the prototype is sourced from a region of high nuclear pleomorphism and this is found to be similar to the lobular region in the input image. Specifically, the highest similarity is to a dense cell region in the top left of the input image. This is somewhat similar to invasive carcinoma as cells are generally more compact and less structured within high grade tumours. However, the similarity is visually weak compared with the benign class prototype, hence the model correctly classifies the image as benign tissue.

Comparisons between the best performing prototypical part network and other methods on the BACH dataset can be seen in table 4.4. Prototypical part networks perform similarly to other methods that make use of a CNN backbone and classifier [5, 152] while providing greater interpretability. Li et al. [89] also use a similar structure but achieved superior performance due to the deep model used (ResNet-50) compared to the ResNet-18 used in this work. Other methods also achieved superior performance to prototypical part networks by utilising model ensembles, multiple resolutions and attention mechanisms. Overall, prototypical-part networks perform close to similar methods while providing greater interpretability.

Table 4.5: Comparison between baselines and ProtoPNet with ResNet-18 backbone on the grading dataset. Validation and Test AUC and accuracy (ACC) ± standard deviation are presented across 5-folds

| Model | pooling | sim metric | Val AUC | Test AUC | Val ACC | Test ACC |
|---|---|---|---|---|---|---|
| Baseline | N/A | N/A | **0.992±0.003** | 0.998±0.001 | 0.925±0.022 | 0.974±0.011 |
| ProtoPnet | avg | Euclidean | 0.984±0.013 | **0.999±0.001** | **0.937±0.020** | 0.986±0.008 |
| ProtoPnet | max | Euclidean | 0.944±0.021 | 0.984±0.009 | 0.889±0.036 | 0.957±0.021 |
| ProtoPnet | avg | cosine | 0.992±0.005 | **0.999±0.001** | 0.936±0.022 | **0.986±0.005** |
| ProtoPnet | max | cosine | 0.961±0.024 | 0.985±0.003 | 0.857±0.055 | 0.906±0.039 |

### 4.5.2 Grading

Model performance on the Grading dataset can be seen in Table 4.5 and Table 4.6 for the ResNet-18 and VGG11 backbones respectively. The results show that the use of the prototypical part network with average pooling and either distance metric outperforms both baselines methods with only the baseline VGG11 model achieving a slightly higher test accuracy. The best performing ResNet-18 model used average pooling and the cosine similarity metric. This configuration achieved a validation AUC of $0.992 \pm 0.005$ and test AUC of $0.999 \pm 0.001$. The best performing VGG11 model also used both average pooling and cosine distance metric. Here the model achieved a validation AUC of $0.999 \pm 0.001$ and test AUC of $0.999 \pm 0.001$. Interestingly, for the VGG-11 backbone, max pooling and cosine similarity metric achieved a higher accuracy on the test set of $0.972 \pm 0.010$. However, the ResNet-18 backbone with average pooling, did achieve the overall highest accuracy of $0.986 \pm 0.005$. The results again confirm that the use of average pooling over max pooling is beneficial for tasks in histopathology as models using average pooling outperformed their max pooling counterparts on average. Additionally, the ResNet-18 backbone benefited more from the change in pooling operator than the VGG-11 backbone. This can be seen from the fact that the VGG-11 with max pooling achieved a higher test set accuracy when compared to average pooling.

The optimal values for $\lambda_4$ for breast cancer grading can be seen in Table 4.7. These values were selected based on the highest average validation AUC across all 5 folds.

Fig 4.9 shows the confusion matrices on the test set for a model trained on each fold. The

Table 4.6: Comparison between baselines and ProtoPNet with VGG11 backbone on the grading dataset. Validation and Test AUC and accuracy (ACC) ± standard deviation are presented across 5-folds

| Model | pooling | sim metric | Val AUC | Test AUC | Val ACC | Test ACC |
|---|---|---|---|---|---|---|
| Baseline | N/A | N/A | 0.988 ± 0.002 | 0.998 ± 0.001 | 0.915 ± 0.023 | **0.974 ± 0.009** |
| ProtoPnet | avg | Euclidean | 0.985±0.013 | 0.998±0.001 | 0.908±0.028 | 0.959±0.016 |
| ProtoPnet | max | Euclidean | 0.968±0.018 | 0.999±0.003 | 0.881±0.036 | 0.919±0.050 |
| ProtoPnet | avg | cosine | **0.999±0.001** | **0.999±0.001** | **0.928±0.024** | 0.966±0.004 |
| ProtoPnet | max | cosine | 0.980±0.006 | 0.997±0.001 | 0.909±0.013 | 0.972±0.010 |

Table 4.7: Orthogonality loss coefficients ($\lambda_4$) for Grading dataset

| config | ResNet-18 | VGG11 |
|---|---|---|
| avg + Euclidean | 0.2 | 0.2 |
| max + Euclidean | 0.6 | 0.4 |
| avg + cosine | 0.4 | 0.6 |
| max + cosine | 0.6 | 0.8 |

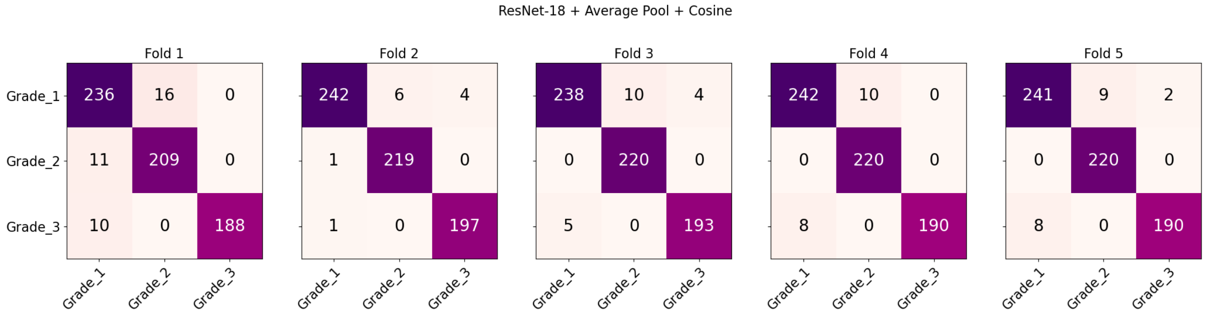


Figure 4.9: Confusion Matrices on the grading test set for a ProtoPNet with ResNet-18 backbone, average pooling and cosine similarity metric trained from each fold.

models all had the ResNet-18 backbone with average pooling and cosine similarity metric. Across all folds, grade 2 images had the highest average recall of 0.989, grade 3 had a recall of 0.968 and grade 1 had a recall of 0.951. These results suggest the model misses fewer moderate and high grade cases compared to lower grade meaning fewer serious cases are missed. Additionally, the grade 3 class had the highest precision of 0.99 resulting in very few false positive grade 3 cases. The grade 1 had a precision of 0.965 and the grade 2 class had a precision of 0.955. The lower precision for the grade 2 is not surprising as it likely the most difficult class to distinguish given that it will share features from both low and high grade tumours. The confusion matrices do highlight however, across all folds, the few grade 3 cases which are misclassified by the model are always predicted to be grade 1. This is an undesirable outcome and should be addressed before application in a real world scenario.

Example model explanations can be seen in Figs 4.10 and 4.11. Fig 4.10 shows a model explanation for a test image with a grade of 1. The three most similar regions to the test image are all from the grade 1 class. Each prototype focuses on areas of tubule formation which is an indicator for a lower grade. Fig 4.11 shows a model explanation for a test image with a grade of 3. The three most similar prototypes are all from regions of dense, poorly differentiated tumour cells with high counts of mitotic figures. These prototypes are extremely similar to the input test image, especially with the right hand side of the image where the tumour cells are the most dense.

Performance of prototypical-part networks on the grading dataset compared to other meth-

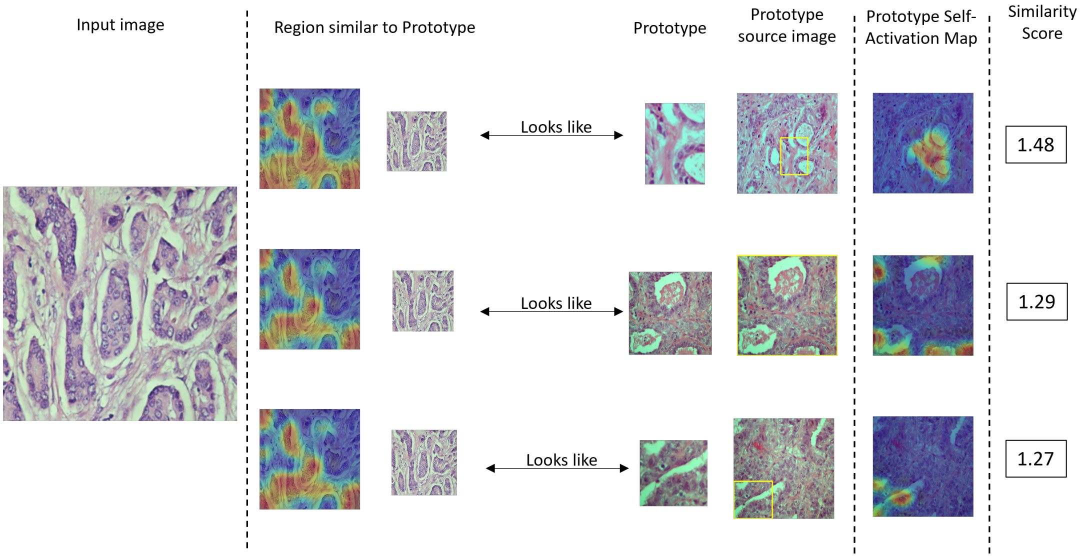


Figure 4.10: Explanation for model prediction of Grade 1 showing the three most similar prototypes to the test image. From left to right, the first image is the input image, the next is a heatmap showing the image region most similar to a prototype, the third is a crop from this heatmap, the fourth is a crop which represents the prototype, the first image is the training image the prototype was taken from, the final image is the activate map of the prototype across the source training image. The similarity score between each prototype and the input image is shown to the far right. The most similar prototype to the input image is at the top of the figure and the third most similar is at the bottom.

Table 4.8: Comparison to other methods on the Grading dataset

| Paper | Method | Test AUC | Test ACC |
|---|---|---|---|
| Dimitropoulos et al. (2017) [47] | Grassmannian VLAD encoding | - | 0.958 |
| Nannia et al. (2018) [103] | CNN Ensemble | - | 0.960 |
| Voon et al. (2022) [144] | CNN | - | 0.952 |
| Ours | ProtoPNet | $0.999 \pm 0.001$ | $0.986 \pm 0.005$ |

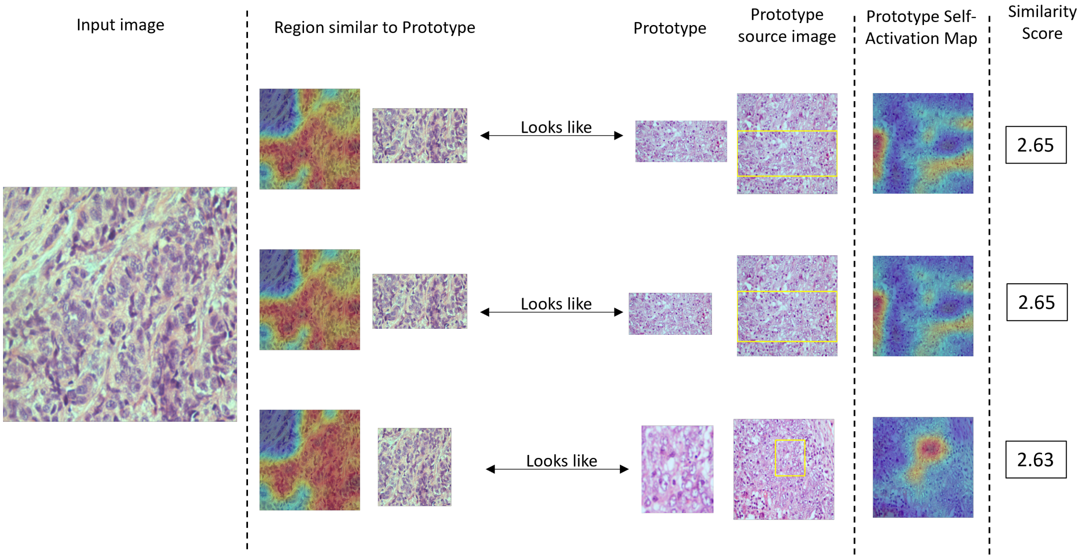


Figure 4.11: Explanation for model prediction of Grade 3 showing the three most similar prototypes to the test image. From left to right, the first image is the input image, the next is a heatmap showing the image region most similar to a prototype, the third is a crop from this heatmap, the fourth is a crop which represents the prototype, the first image is the training image the prototype was taken from, the final image is the activate map of the prototype across the source training image. The similarity score between each prototype and the input image is shown to the far right. The most similar prototype to the input image is at the top of the figure and the third most similar is at the bottom.

ods is shown in table 4.8. The results show that the prototypical-part networks outperforms other methods on the grading dataset. On top of this, the model provided greater interpretabiltiy compared to these methods making it a viable model choice for ROI grading tasks.

### 4.5.3 Orthogonality Loss

To investigate the benefit of the orthogonality loss on prototype discovery the prototypes from models trained with and without the loss term can be compared. Fig A.1 in Appendix A shows the 40 prototypes discovered with and without the orthogonality loss on the BACH dataset. The prototypes are taken from models trained using the ResNet-18 backbone with average pooling and cosine distance similarity metric on the first fold of the training and the validation set which produced the highest AUC. The inclusion of the orthogonality loss clearly produces a greater diversity of prototypical parts most notably for benign, in-situ and invasive tissue images. Diverse ranges of prototypes were discovered for normal images with and without the orthogonality loss. Fig A.2 in Appendix A shows the 30 prototypes discovered with and without the orthogonality loss on the Grading dataset. Again, the prototypes are taken from models trained using the ResNet-18 backbone with average pooling and cosine distance similarity metric on the first fold of the training and validation set which produced the highest AUC. The figure shows that the orthogonality loss slightly increases the prototype diversity for the grade 2 class however the loss term has had little impact on the grade 1 and grade 3 prototypes. Overall, the orthogonality loss did not increase prototype diversity for the breast cancer grading in comparison to breast cancer sub-typing. This could be due to the grading dataset having less intra-class diversity compared to the BACH dataset. As the grading set images are taken at a higher magnification it is likely that they capture images with similar morphological appearances compared with images taken at a lower magnification.

Fig 4.12 shows the effect of increasing the orthogonality loss coefficient $\lambda_4$ on the test set AUC and accuracy on average across all training folds. The results on the BACH dataset show that, for ResNet-18 there is a slight increase in test AUC and test accuracy with orthogonality loss. Increasing $\lambda_4$ also appears to improve performance for the VGG11 backbone up to a value of 0.8, after this performance decreases when using the Euclidean distance similarity metric.

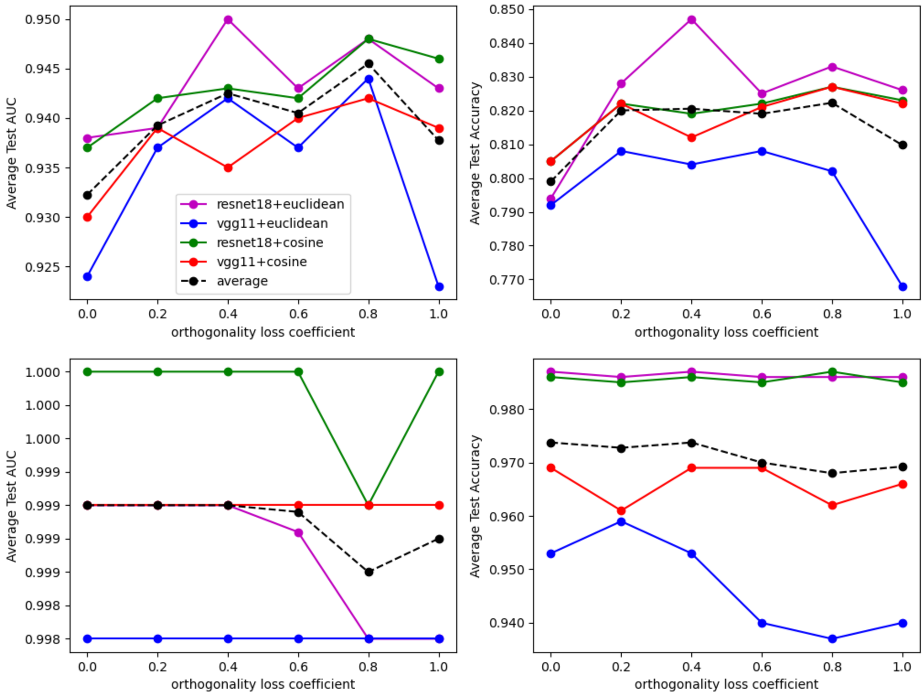


Figure 4.12: Change in average test AUC (left) and average test accuracy (right) across all folds with increasing orthogonality loss coefficient for ResNet-18 and VGG11 backbones with average pooling. Top: shows the results on the BACH dataset. Bottom: shows results on the Grading dataset.

On the other hand, the VGG11 backbone with cosine distance metric performance does not suffer to the same extent when $\lambda_4$ is increased past 0.8. Model performance on the grading dataset does decrease on average with increased $\lambda_4$. The ResNet-18 backbone with Euclidean distance metric saw the greatest decrease in test AUC with increased $\lambda_4$ while the VGG-11 with Euclidean distance metric saw the greatest decrease in test accuracy with increased $\lambda_4$. Overall, the benefit of including the orthogonality loss is increased prototype diversity leading to more sensible and appealing model explanations. The changes in model performance are negligible for both datasets.

### 4.5.4 Pooling function

The choice of pooling function appears to impact model performance with average pooling outperforming max pooling in all model configurations. However, the choice of pooling also affects the prototype similarity maps and thus what each prototype represents. This additionally changes the appearance of the prototypes when they are used for explanations. Fig A.3 shows the difference in the similarity heatmaps between a ResNet-18 model trained with average pooling on the top and max pooling on the bottom. The figure clearly shows that the average pooling approach produces broader heatmaps which can highlight whole lobules, necrotic tissue regions and even whole regions of tumour. On the other hand, the max pooling approach highlights portions of lobules or even specific ducts as well as small clusters of tumour cells. This is likely to be the defining reason behind the difference in model performance. The original prototypical-part-network was designed and tested on natural images such as those from the CUB-200-2011 dataset (https://data.caltech.edu/records/65de6-vp158). The dataset contains images of birds which often have key components that appear once or twice in an image. For example, a single beak or two wings. However, histopathology images can contain vast regions of many tumour cells and ductal structures. In addition, tumour cells could be found on opposing corners of a histopathology image whereas the features or a bird are often contained in a similar regions within an image. Therefore a pooling function that takes into account these factors is likely to achieve superior performance.

### 4.5.5 Number of Prototypes

To understand how the number of prototypes impacts model performance and prototype discovery the separate models were trained with different numbers of prototypes per class. The predetermined optimal value for $\lambda_4$ was used depending on the CNN backbone architecture and similarity metric employed. Fig 4.13 shows how model performance changed with different quantities of prototypes with equal numbers in each class. The results show that, for a ProtoPNet with ResNet-18 backbone, average pooling and cosine similarity metric the best performance is achieved with 32 prototypes (8 per class). Using 40 prototypes (10 per class) achieved the second best performance. Overall, the number of prototypes had little overall

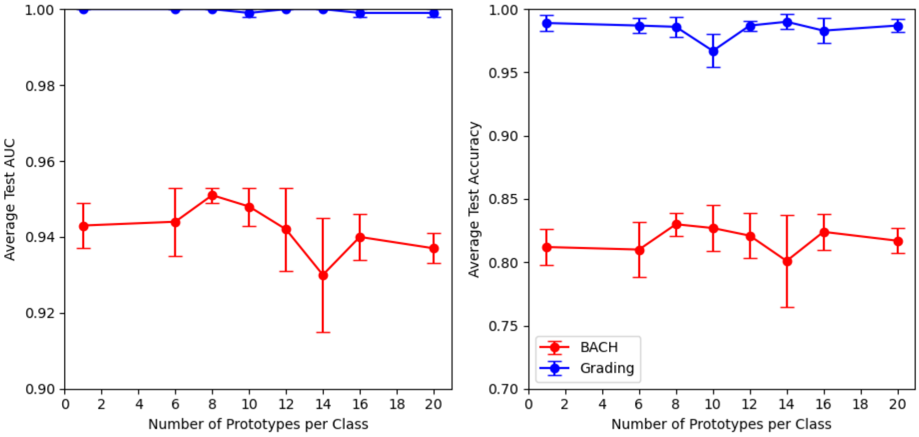


Figure 4.13: Change in test AUC and test accuracy with change in number of prototypes per class. Graphs show both BACH and Grading dataset results.

impact on model performance with the best and worse performing models having a difference in test AUC of 0.021 and test accuracy of 0.029.

### 4.5.6 Failure Cases

There are several occasions where the model under-performs and predicts the class incorrectly. Identifying these cases is particularly important especially in a healthcare setting, where an incorrect diagnosis can result in sub-optimal treatment for patients. Fig 4.14 shows four examples where the prototypical part network incorrectly predicts the breast cancer sub-type.

In image a) the image contains in-situ carcinoma however the model predicts Benign incorrectly. From the prototypes and heatmaps it can be seen that the necrotic region within the duct has a similar appearance to the central area of a benign duct. Although the images have different colours, both regions are highly homogeneous which might have caused the false diagnosis. From the model explanation it is likely that a pathologist would not trust the model prediction as it is clear the prototypes are focusing on the wrong image region.

In image b) the image contains Benign tissue however the model predicts this as carcinoma in-situ. The most similar prototype to the input benign image is a prototype of in-situ carcinoma representing ductal cells. The model incorrectly compares this with the necrotic tissue within the benign duct. The second most similar prototype is from the Normal tissue class and focuses on the stroma surrounding the duct. Finally, the third most similar prototype is from the invasive class and again focuses on the necrotic tissue within the duct. The fact that the top three prototypes are all from separate classes indicates that the model is uncertain in its prediction and should therefore be reviewed by a pathologist.

In image c) the image contains invasive carcinoma however the model predicts this as in-situ. This is a particularly hard case as the input image does resemble a ductal structure where the tumour cells appear contained. This is reflected in the fact that the three model similar prototypes are all from the in-situ class and are all found similar to the ductal structure on the right hand side of the image. In this particular case it would be hard to flag this as model uncertainty.

In image d) this image contains in-situ carcinoma however the model predicts this normal. This is a particularly bad misdiagnosis. The most similar prototype contains an image of a duct within a region of adipose. This is found to be similar to the input image due to the presence of a single duct near to fat tissue. The second most similar prototype is of a healthy duct which is found to be similar to the central duct within the input image. Finally, the third prototype activates highly on regions of stroma surrounding the duct. From the BACH dataset images of normal tissue appear to contain larger regions of stroma in comparison to images from the in-situ class. The input image could be considered an outlier for this reason and hence is misclassified by the model. Because of this spurious reasoning a pathologist might pick up on this misclassification for review.

For breast cancer grading, as mentioned in section 4.5.2, the biggest pitfall were the few cases where grade 3 inputs were classified as grade 1. An example of a misclassified grade 3 case can be seen in fig 4.15. The three most similar prototypes in order of similarity are a grade 1, grade 3 and a grade 1 prototype. This suggests that the model does find the input image similar to both a grade 1 and grade 3 cases which could flag for an end user that the image is difficult for the model to predict. The prototypes for grade 1 appear to be similar to an image region in the top right of the image. Here there are fewer tumour cells compared to the rest of the input image. The few tumour cells that are present are compared to smaller lobule structures found within lower grade tumours. The grade 3 prototypes appears to activate more homogeneously

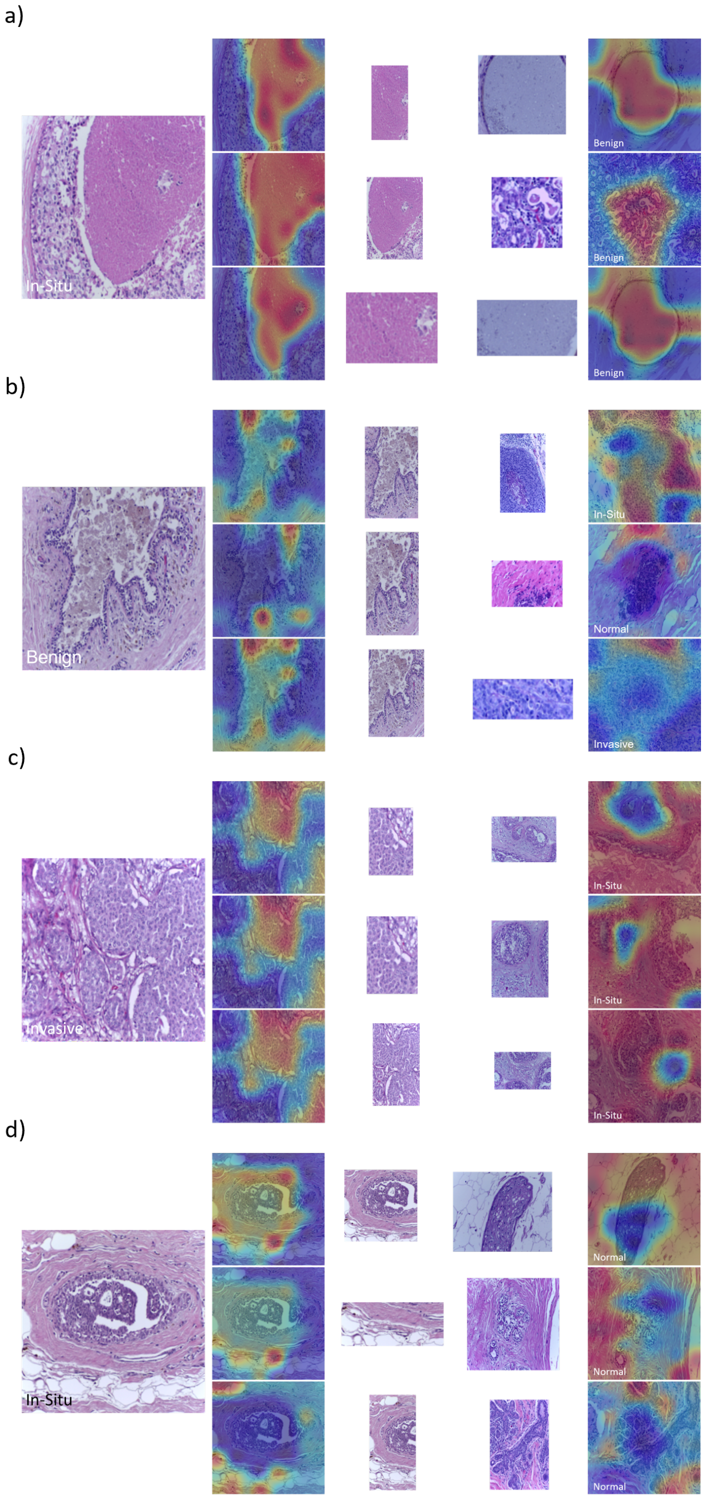


Figure 4.14: Failure cases for a prototypical part network with a ResNet-18 backbone, average pooling and cosine distance similarity metric. Each column from left to right represents the test image, prototype similarity heatmap with test image, extracted image region from heat map, the prototype similar to the region within the input image, the prototype similarity map with it's source image. Each row (1-3) represents the prototypes with the greatest similarity to the test image. The top row is the most similar prototype while the bottom row is the least similar. a) Failure case for in-situ carcinoma. b) Failure case for Benign tissue. c) Failure case for invasive carcinoma. d) Failure case for in-situ carcinoma

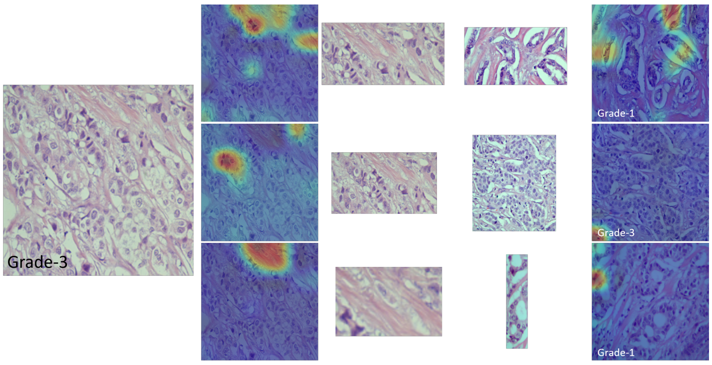


Figure 4.15: Failure cases for prediciton of grade-3 using a prototypical part network with a ResNet-18 backbone, average pooling and cosine distance similarity metric. Each column from left to right represents the test image, prototype similarity heatmap with test image, extracted image region from heat map, the prototype similar to the region within the input image, the prototype similarity map with it's source image.

across the rest of the image with particular focus on one left central region. In these regions there are more densely packed tumour cells indicating a higher grade cancer. This result shows that grade 3 images with small regions of low density tumour cells can be misclassified for grade 1. However, there are very few cases where grade 3 inputs are misclassified, suggesting that this scenario is an outlier case.

## 4.6 Conclusion

In conclusion, the Prototypical Part Network has been evaluated in the context of digital pathology slide analysis and has been found to outperform the black-box counterparts. This performance increase could be because the Prototypical Part Networks use more parameters than the baseline methods, due to the inclusion of the prototype layers. However, the baseline models achieved superior performance when the Prototypical Part Network used max pooling of the similarity scores. Therefore, the greater performance is likely a combination of the increased number of parameters and the appropriate selection of pooling operator. Making use of average pooling allows the similarity of the prototypes to be considered across all regions of the image, increased model performance. This is likely important for histopathology slide analysis as tumour regions can extend across different regions of an input image.

The methods were evaluated on two open source ROI classification tasks. The first was breast cancer sub-typing while the second was breast cancer grading. Several architectural changes were made and evaluated and were found to improve performance. As mentioned, one change made was the inclusion of average pooling over max pooling, this achieved the greatest performance increase. Another change was the use of an orthogonal loss term to force the prototype vectors to be orthogonal to one another and aimed to increase prototype diversity. the final change was to use a cosine similarity loss over Euclidean distance which did perform slightly better than the Euclidean distance metric. Two baseline CNN models were also tested, these were the ResNet-18 and VGG-11.

There are several limitations of the prototypical part network for histopathology slide analysis. The first is that histopathology slides contain very varied morphologies and the method found it difficult to find prototypes that generalise well to all morphologies which can produce undesirable explanations. To create better explanations, the use of even more prototypes could be explored but this could result in over fitting and longer training times. Alternatively, each prototype could be represented by a small ensemble of prototypes or a small cluster of prototypes which collectively represent the same prototypical part but each capture a slight variation in morphology. Another limitation is the scale of the datasets. ROIs provide less diagnostic information than WSI and therefore make the method less valuable than those that can analyse WSI. Future work could look to incorporate the idea of prototypes into methods for WSI analysis.

Chapter 5

# Weakly-Supervised Learning for NPI prediction

## 5.1 Introduction

Prognostic factors in breast cancer analysis play an important role in determining the optimal treatment pathways for patients and allow for identification of groups who would or would not benefit from additional treatment. Effective prognostic factors allow for wide separation of outcomes between different groups with adequate numbers in each group. As introduced in 2.2.5, the NPI is one such prognostic factor used in breast cancer analysis. It is determined using three other prognostic factors each found to be independently associated with survival on multivariate analysis [53]. These three factors are: 1) The size of the greatest dimension of the invasive tumour, 2) The cancer grade, 3) The axillary lymph node stage. The NPI is calculated using the formula NPI = Grade (1 to 3) + Node (1 to 3) + (size of invasive carcinoma in cm × 0.2) and patients have been stratified using the NPI into six groups with good numbers in each group and very divergent outcomes [78], which can be seen in Table 2.1 in chapter 2. The NPI is important for determining the most appropriate treatment for a patient, including whether or not chemotherapy or radiation therapy is necessary. The NPI is also a useful tool for predicting the likelihood of recurrence and survival, which can help doctors and patients make informed decisions about ongoing care and follow-up. Additionally, the NPI is widely used in clinical trials and research studies, as it provides a standardised method for stratifying patients based on their prognosis. Overall, the NPI is an important tool for guiding clinical decision-making, improving patient outcomes, and advancing our understanding of breast cancer prognosis and treatment.

In this chapter, weakly-supervised approaches to the prediction of the NPI group [78] from histopathology images are explored. Specifically two approaches are investigated. The first approach aims to predict each prognostic factor individually and then combine them together to predict the NPI group. This approach will be referred to as the *part prediction* approach and aims to mimic the pathologist workflow. The second approach directly predicts the NPI group by combining both primary tumour and lymph node images through a single model. The approach is trained directly on prognostic group prediction and thus aims to assess if morphological features within the WSIs are a predictor of prognostic group. This will be referred to as the *direct prediction* approach.

Each approach makes use of Multiple Instance Learning (MIL), and more specifically, Attention-based MIL (ABMIL) [67] with multi-class attention pooling [98] to produce slide embeddings. In this framework the WSI is referred to as a bag which is comprised of multiple instances. The instances are patches extracted from non-background regions of the WSI. The use of attention pooling weights the contribution of each instance (patch) to the final bag embedding. This further allows inspection of discriminative instances. In the part prediction approach a separate model is trained for grade prediction from the primary site and for lymph node stage prediction from the lymph node. The size of the primary tumour was taken from the pathologist reports associated with the patients. This is because the primary tumour can extend over multiple slides and therefore in many cases cannot be measured from a single slide. For direct prediction of the NPI the WSI embeddings between slides were fused. Fusion methods explored were; vector concatenation, average pooling, vector addition, the Hadamard product and a self-attention module [142].

The main contributions of this chapter are: 1) Two frameworks for prediction of NPI prognostic group based on weakly supervised MIL are developed, the first predicts the individual NPI score components while the second directly predicts the NPI group, 2) The use of several fusion operators, including vector concatenation, average pooling, vector addition, Hadamard product and a self-attention head to aggregate features from the primary breast site and lymph node are investigated, 3) It is found that the part prediction approach outperforms direct group prediction.

## 5.2 Related Works

There are no other works which aim at predicting the NPI from WSI. In addition, there are no works which incorporate features from both the primary breast tumour and lymph nodes in an attempt to improve performance. There are however many works which predict the breast cancer grade, which is a core component of the NPI as well as works which look at patient survival analysis which NPI aims to predict. Additionally, there are several works which combine slides with different staining as well as multi-model data and are therefore related to this work through the fusion methods explored. This work is also related to the range of multiple instance learning methodologies applied to WSI which have already been covered in section 2.5.2.

### 5.2.1 Breast Cancer Grading

An early work for automated grading of breast cancer used spectral clustering with textural (Gabor, Grey Level and Haralick) and graph (Voronoi diagram, Delaunay triangulation, minimum spanning tree, nuclear characteristics) features on a small sample of 48 breast biopsies [49]. Additionally, a multi-view framework was also introduced to extract textural and graph-based cell features from different image magnifications [11]. One limitation of these approaches was the required identification of histologic primitives such as nuclei. To address this, other approaches attempted to model the data directly. Dimitropoulos et al [47] did this through modelling the breast cancer histological images as a set of spatially evolving multidimensional signals, which are mapped and encoded on the Grassmann manifold. So far, the mentioned approaches have only used small ROIs. This is a limiting factor as these ROIs need to be identified and cropped from an image. Methods that could use a patient's WSI without this pre-processing step would be beneficial. To initially address this, Couture et al used Tumour

Microarray (TMA) images for breast cancer grading [40]. These images can be thousands of pixels in both height and width making a prediction task harder. To solve this, a probabilistic model can be formed to show how likely each image region is to belong to each class, with these probabilities aggregated across all image regions to form a prediction for the tumour as a whole. However, the main criticism of TMA images is that the amount of tissue analysed is limited and may not be representative of the whole specimen. Although they provide more context that ROI crops, the gold standard would be WSI. For analysis of WSI, Wetstein et al [154] used a MIL method with RNN aggregation [20] for breast cancer grading. They also propose a multi-task framework, to not only predict the individual components of grade (tubule formation, mitotic count and nuclear atypia), but also the HER2 status, aiming to enrich the feature space. The method was able to classify between low/moderate grade cases and high grade cases with an accuracy of 77%. This increased to 79% using the multi-task framework. One limitation of this work was the binary classification approach. Both low and moderate cases were grouped together into a single class due to dataset restrictions. The work by Wang et al [151] addresses this with an ensemble of CNN models for prediction of high and low grade patches from tumour regions within a WSI. Additionally, they re-stratify moderate grade groups into moderate low and moderate high. However, this method required ground-truth, patch level annotations of tumour regions whereas the MIL RNN approach did not.

### 5.2.2 Survival Analysis

Yao et al [163] first clustered input feature embeddings which generates several distinguished phenotype groups. The patches from each group are fed through a Siamese FCN which produces a single embedding vector per group using a final pooling layer. These embeddings are then passed through an attention based MIL module to get the final slide representation which is ultimately used for survival prediction. Abbet et al [1] also utilised a clustering based approach, this time through clustering the embedding space from an auto-encoder using spherical k-means. Patches were assigned to a dictionary of cluster centres and features were extracted based on the probability of patch clusters and local interactions between clusters. Wulczyn et al [155] changed the survival regression problem into a classification problem by discretising survival time into survival intervals. Liu et al [92] created six-channel images; RGB colour channels plus nucleus segmentation maps, tumour infiltrating lymphocyte (TIL) maps and tumour maps as additional channels. They similarly convert the regression problem into a classification one. Courtiol et al [39] proposed MescoNet for mesothelioma tumour survival prediction which uses a 1D convolutional layer to predict survival scores for image patches. The top 10 and bottom 10 scores are then input to a MLP for survival prediction. Li et al [87] applied graph neural networks to task of survival prediction. Patches were considered the nodes of a graph and edges were connected between patches based on a thresholded euclidean distance between patch features. Patches are also randomly sampled across the slide. Chen et al [31] proposed a patch-based graph convolutional network for survival prediction by also connecting nodes due to their spatial arrangement in the slide. Additionally, all patches per slide are utilised. Wetstein et al [154] used the RNN MIL method proposed by Campanella et al [20] for breast cancer survival prediction and introduced grade and HER2 prediction in a multi-task framework to provide extra discriminatory information to the model. Shaban et al [120] calculated a score from regions of tumour associated stroma and infiltrating lymphocytes to predict survival in head and neck squamous cell carcinoma.

### 5.2.3 Data Fusion Approaches

Data fusion approaches assess the benefit of combining data from different modalities. The intuition being that one modality might provide information which cannot be obtained from the other and vice versa. Therefore, when combined, the predictive capabilities of a trained model should exceed a model using only a single data stream. Combining slides with different stains has been shown to be a successful approach. For example, Graph Neural Networks were fit to slides stained with both H&E and Trichrome (TC) and the graph outputs were combined together for medical evaluation of nonalcoholic steatohepatitis [50]. Cell-graphs have also been fit to IHC datasets containing slides stained with Ki67, CK20, P16 and CD20 which were fused together using a self-attention module [102]. Genomic features have also been combined with WSI through a co-attention transformer framework for survival prediction [29, 80].

## 5.3 Methods

### 5.3.1 Overview

For predicting breast cancer grade and lymph node stage, attention-based MIL is used which weights the contribution of each patch within a WSI towards the final classification. In the part prediction approach a separate MIL model is trained for each classification task. In the direct approach a single attention based MIL model is trained to process both the primary breast tumour and lymph node information thus producing multiple slide embeddings whose information should then be aggregated together to produce a joint slide embedding. The specifics of MIL with attention based pooling will now be introduced followed by an explanation of the way it is used for both grade, lymph node stage and direct group learning.

### 5.3.2 Attention-based MIL pooling

In order to produce an embedding for a WSI, attention-based MIL pooling (ABMIL) [67] is used due to its good performance and interpretability from viewing the attention weights. This method builds upon the standard multiple instance assumption introduced in section 2.5.2. First, a WSI is split into $L$ non-overlapping patches where each patch $x_l$ is of size $256 \times 256$ pixels. Therefore, the WSI is represented as a set of patches $X = \{x_1, ..., x_L\}$. A neural network $f$ is used as a feature extractor and produces an embedding $z_l = f(x_l) \in \mathbb{R}^{b \times 1}$ for each image in a bag $X = \{z_1, ..., z_L\}$ where $b$ is the length of the embedding vector. The instance embeddings for each bag are then passed to an attention block for attention pooling. Here a fully connected layer compresses each $b$ dimensional embedding to a 512-dimensional vector $h_l$. Then for each 512-dimensional vector an attention score is produced such that the bag embedding is given by

$$B = \sum_{l=1}^{L} a_l h_l \tag{5.1}$$

where:

$$a_l = \frac{exp\{W_a(tanh(Vh_l^\top) \odot sig(Uh_l^\top))\}}{\sum_{j=1}^{L} exp\{W_a(tanh(Vh_j^\top) \odot sig(Uh_j^\top))\}} \tag{5.2}$$

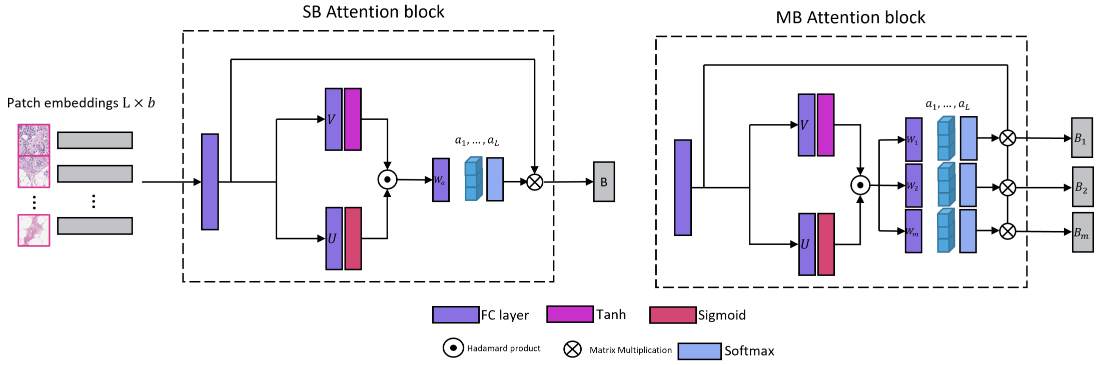


Figure 5.1: Architecture of the gated-attention block [67]. Left single branch attention, Right multi-branch attention. Patch embeddings from a WSI are input to an initial fully-connected layer (FC) to compress to $L \times 512$. Then the compressed embeddings are input to FC layers $V$ and $U$ which are passed through a hyperbolic tangent and sigmoid function respectively. The outputs are fused by a Hadamard product and passed through a final FC layer $W_a$ to output $L$ attention coefficients. In the case of multi-branch attention there are $n$ FC layers one for each of $n$ classes. The attention coefficients are then normalised via a softmax function before being matrix multiplied by the compressed patch embeddings. This produces a single bag embedding $B \in \mathbb{R}^{1\times512}$ for the slide or $B \in \mathbb{R}^{n\times512}$
in the case of multi-branch attention.

$U \in \mathbb{R}^{256\times512}$, $V \in \mathbb{R}^{256\times512}$ and $W_a \in \mathbb{R}^{1\times256}$ are weight matrices, $\odot$ is an element-wise multiplication, $sig(.)$ is the sigmoid function and $tanh(.)$ is the hyperbolic tangent. The left of Fig 5.1 shows the architecture for the attention block.

This is often referred to as single branch (SB) attention as a single attention value is given to each instance. SB attention is commonly adopted for binary classification tasks where only the attention for the positive class is of interest. In the case of multi-class problems, where the positive information for each class is of interest multi-branch (MB) is introduced. In MB attention (eq 5.3) [98] $U \in \mathbb{R}^{256\times512}$ and $V \in \mathbb{R}^{256\times512}$ are still shared across all classes however $W_{a,1}, ..., W_{a,n} \in \mathbb{R}^{1\times256}$ are instead $n$ parallel branches for each of the $n$ classes to be classified. This treats the multi-class problem as $n$ binary classification tasks, one for each class. For each binary classification an attention value is produced for each instance which indicates the relevance of that instance to a particular class. From the MB attention $n$ 512-dimensional embeddings are produced, one for each class. $n$ classifiers $W_{c,1}, ..., W_{c,n} \in \mathbb{R}^{1\times512}$ then map each embedding to an unnormalised slide level score for each class as shown to the right in Fig 5.1. In contrast, for the SB attention a single 512-dimensional embedding is produced and a classifier $W_c \in \mathbb{R}^{n\times 512}$ outputs unnormalised scores for each class. Regardless of the attention used the slide scores are normalised using a softmax function to return the probability of each class for a particular WSI.

$$a_{m,l} = \frac{exp\{W_{a,m}(tanh(Vh_l^\top) \odot sig(Uh_l^\top))\}}{\sum_{j=1}^{L} exp\{W_{a,m}(tanh(Vh_j^\top) \odot sig(Uh_j^\top))\}} \tag{5.3}$$

### 5.3.3 Part Prediction

Using the ABMIL approach with MB attention, separate models were trained to predict breast cancer grade from the primary site and lymph node stage from the lymph node. This is the part prediction approach and the full architecture can be seen in Fig 5.2. The patch extraction and feature embedding steps are shared between the grading model and the staging model. However, the Attention-Based Aggregation and MB classification sections are trained separately for each task (grading and staging). The Grading Attention module is the attention model trained for breast cancer grade prediction while the Staging Attention module is trained for lymph node staging. The weights of these models are updated independently, meaning that the performance of one is not considered when training the other.

After training of the separate grade and lymph node staging models, the predictions were then combined with the tumour size from the pathologist reports to return the NPI and stratify patients into groups.

### 5.3.4 Direct Prediction and Fusion Methods

For the direct prediction approach, both the breast primary tumour and a corresponding lymph node, a bag embedding is produced for each slide in the case of SB attention and one embedding per class for MB attention. The bag embeddings are produced by passing the patches from each slide through the attention module independently. This produces a primary tumour embedding $B_P$ and a lymph node embedding $B_{LN}$. To incorporate information from both slides into the final decision several fusion methods are explored as shown in Fig 5.3.

- Vector concatenation: The two slide embeddings are concatenated together to produce a 1024-dimensional embedding.

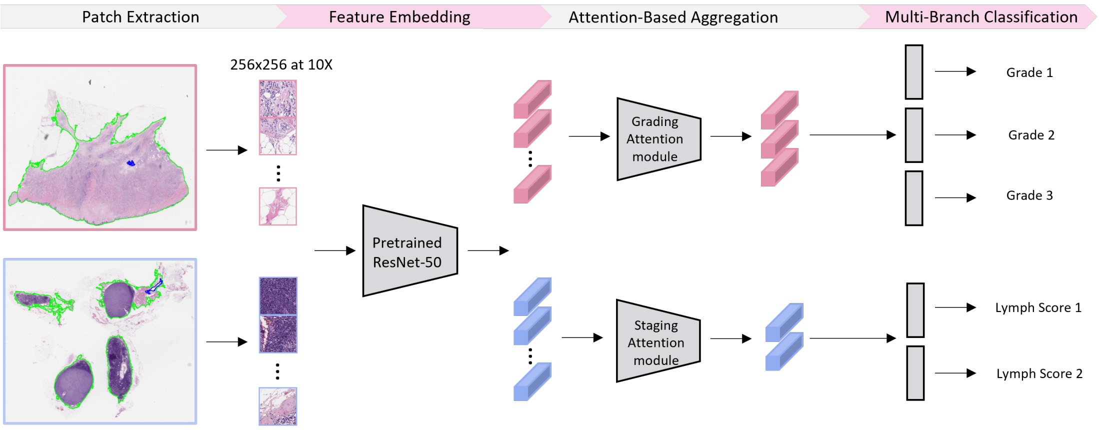


Figure 5.2: Architecture of the part prediction approach methodology. Primary site and lymph node WSI are first split into non-overlapping patches of size $256 \times 256$ pixels. Patches from each slide are passed through separate MB attention blocks which have been outlined in section 5.3.2 The primary site embeddings are used for grade prediction while the lymph node embeddings are for lymph node stage prediction. The outputs are combined with the pathologist reported Invasive tumour size to predict the final NPI group.

– Average-pool: The average of the two slide embeddings is taken, producing a 512-dimensional embedding.

– Vector Addition: Performs vector addition of slide embeddings after projecting them into a common latent space.

$$B' = W_P B_P + W_{LN} B_{LN} \tag{5.4}$$

where $W_T$ and $W_{LN}$ are learnt FC layers.

– Hadamard Product: The Hadamard product of the slide embeddings after projecting them into a common latent space.

$$B' = (W_P B_P \odot W_{LN} B_{LN}) \tag{5.5}$$

where $W_P$ and $W_{LN}$ are learnt FC layers.

– Self-Attention: Attention coefficients are determined for each slide embedding and are dependent on one another. This is further explained in section 5.3.5.

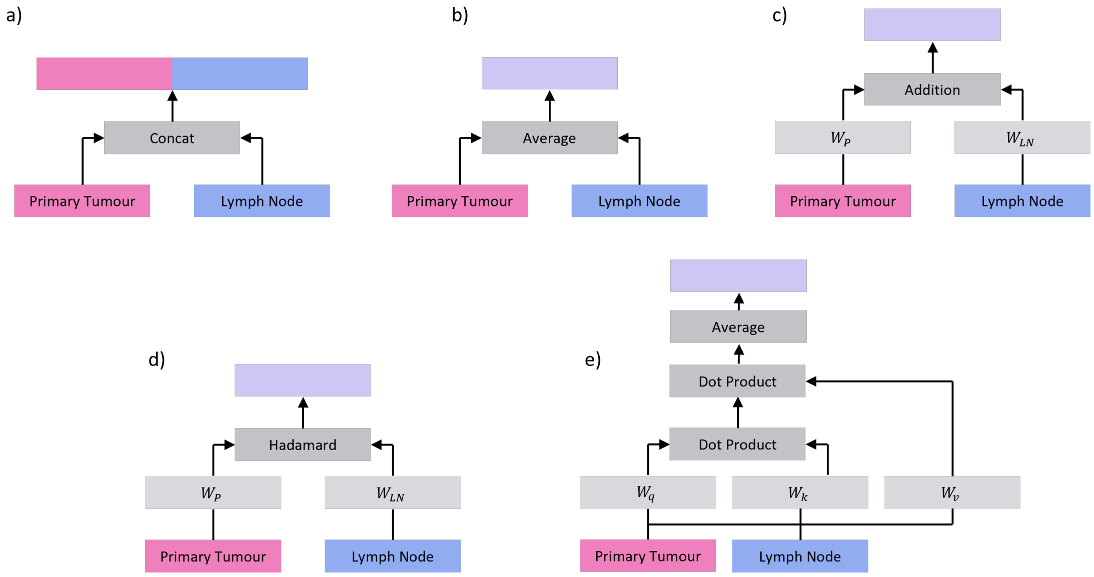


Figure 5.3: Fusion Methods: a) Vector concatenation, b) Average-pool, c) Vector Addition, d) Hadamard Product, e) Self-Attention

### 5.3.5 Self-Attention

Instead of concatenating the WSI embedding for the primary breast site and the corresponding lymph node, a self-attention module can be used to further aggregate the embeddings into a more meaningful one. Unlike the previous attention module, self-attention takes into account the dependencies between the primary site and lymph node. This means that the attention weight applied to either the primary site or lymph node is dependent on both embeddings. The primary tumour $B_P$ and lymph node $B_{LN}$ embeddings are concatenated to $B = \{B_P, B_{LN}\}$. A self-attention layer then maps this to queries, keys and values using weight matrices $W_q, W_k, W_v \in \mathbb{R}^{512\times512}$. The self-attention embeddings for the primary site and lymph node are calculated using

$$B' = softmax(\frac{W_Q B W_K^\top B}{\sqrt{d}}) W_V B \tag{5.6}$$

where $d$ is the length of the embedding vectors. In the case of multi-head attention the query, key and value vectors are further split up depending on the number of heads, with each split being independently passed through a separate head. 8 heads are used which results in the 512-dimensional key, query and value vectors being split into 8 64-dimensional vectors each.

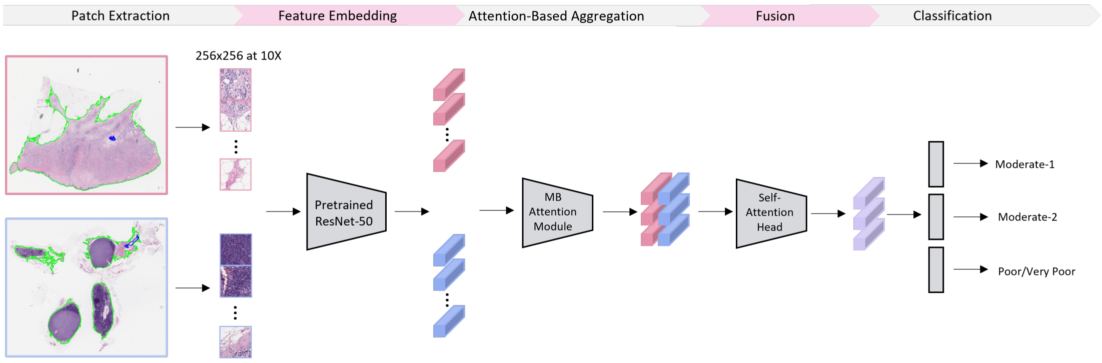


Figure 5.4: Architecture of the direct approach methodology. Primary site and lymph node WSI are first split into non-overlapping patches of size 256 × 256 pixels. Patches from each slide are passed through a MB attention block which have been outlined in section 5.3.2. This produces 3 slide embeddings representing the information required for predicting each class (3 classes). Each pair of slide embeddings for each class are then passed to a fusion head to aggregate embeddings. The 3 aggregated embeddings are then passed to 3 dense layers, one for each class. Each layer predicts the probability of a particular class.

The multi-head approach provides greater power to encode multiple relationships between slide embeddings. Then the self-attention embeddings $B' = \{B'_P, B'_{LN}\}$, are input to a FC layer with weight vector $W \in \mathbb{R}^{512\times512}$. This is then normalised and passed through a MLP with one hidden layer with Gaussian Error Linear Unit (GELU) activation. To predict the NPI the embeddings $B'_P$ and $B'_{LN}$ are averaged to produce a single embedding which is passed to a linear classifier $W_c \in \mathbb{R}^{1\times512}$ and outputs a single value $s_n$ for the predicted index. The full architecture can be seen in Fig 5.4.

### 5.3.6 Interpretability

Similar to other works using attention-based pooling methods for WSI classification [67, 95, 90], inspection of the patch aggregation process provides insights into which patches are the most important to the final WSI embedding. As slide embeddings are produced for each class, the attention weights placed on patches that contributed to each prediction can be viewed. Where the attention weights are higher (closer to 1 after applying softmax to all attention weights on a slide) the patches are more important to a prediction. Therefore, the important regions within a slide can be viewed to determine if the model is using clinically relevant information, such as paying more attention to tumour regions when predicting poor prognosis.

Additionally, the self-attention weights can be viewed to determine the contribution of the primary site and lymph node to the final prediction. This allows further insight into the importance of each WSI.

## 5.4 Experiments

The aim of the experiments is to evaluate the several different approaches to the task of NPI group prediction. The methods tested are 1) The individual prediction of grade and lymph node stage combined with invasive tumour size from pathologists reports to determine grade, 2) Fusing the primary tumour and lymph node images together for direct prediction of NPI. Stratified 5-fold cross validation is used to perform robust analysis of the methodologies.

### 5.4.1 Dataset

The LTHT dataset was used here which consists of 163 primary tumour and lymph node paired WSI resulting in 326 slides in total. The slides were scanned using an Aperio scanner at 0.25 micrometers-per-pixel. Patches were extracted from the foreground tissue regions of each slide to reduce redundant patch extraction from background regions. Non-overlapping patches of size 256 × 256 pixels at 10× magnification were extracted as well as at 20× magnification. These magnifications were chosen to strike a balance between collecting high level detail without having excessive bag sizes. Large bag sizes can result in extended training times as well as noisy bags due to an overload of feature embeddings. On average there were 2954 patches extracted from primary slides and 1243 from lymph node slides at 10× magnification and 11350 from primary slides and 4797 from lymph node slides at 20× magnification. The whole dataset was split into a development and test split in a stratified fashion using the distribution of prognostic group labels. Detailed information about the development and test datasets can be seen in Table 5.1. 80% of the overall data was used for the development set and 20% was used for the test set. The development set was used for stratified 5-fold cross validation. Of the 163 cases, 130 cases were used for training and validation while 33 were used for testing. Due to the class imbalances the Good prognosis class was ignored in all experiments reducing the development set size to 127 cases and the test set size to 32 cases. All of the models were evaluated on two separate experiments, the first was classification between moderate-1, moderate-2 and combined poor and very poor classes. The second experiment was a binary classification task between combined moderate classes and combined poor and very poor classes. Detailed explanations of slide preparation and patch extraction can be found in section 3.

### 5.4.2 Implementation Details

For feature extraction of patch level features a ResNet-50 [60] architecture pretrained on ImageNet is used. The ResNet-50 parameters were frozen for feature extraction. For each patch of size 256 × 256 this produced a 1024-dimensional feature embedding. For training the MIL models Adam optimiser was used with a fixed learning rate of $1 \times 10^{-4}$ and a weight decay of $1 \times 10^{-5}$. These parameters were chosen as they are recommended for other Attention based MIL approaches such as CLAM [98]. Cross-Entropy Loss was used for prediction of NPI group, grade and lymph node stage. For prediction of grade and lymph node stage the cross-entropy loss was weighted for each class due to the class imbalance. The weight for a class was given by $1 - \frac{n_{class}}{n_{total}}$. All models were trained for a maximum of 50 epochs and early stopping was implemented if the validation weighted F1-score did not improve for 10 epochs. 5-fold cross validation was used when training and the model with the highest validation weighted F1-score was selected for testing in each fold. Training was implemented with a single NVIDIA V100 GPU on the ARC4 HPC cluster. The models were implemented in PyTorch [108].

## 5.5 Results and Discussion

### 5.5.1 Cross-validated Model Performance

The model performance on the test set for prediction of moderate-1 vs moderate-2 vs poor/very poor using patches extracted at 10× can be seen in Table 5.2. Note that there is no AUC given for the part prediction approach as this approach combines multiple separate classifications together so it is difficult to adjust for different thresholds. The part prediction approach outperformed all direct prediction methods achieving the highest accuracy of $0.727 \pm 0.064$ and F1-score of $0.730 \pm 0.064$. Furthermore, a test accuracy of $0.612 \pm 0.030$ and test F1-score of

Table 5.1: LTHT patient characteristics in the development and test datasets. A Chi-squared test was used to determine any statistical significance between the categorical variables between datasets and Welch's t-test was used for numerical variables.

| Patient Characteristics | Development dataset | Test dataset | $p$-value |
|---|---|---|---|
| $n$ | 130 | 33 | |
| **Tumour Grade** ($n$ (%)) | | | 0.485 |
| Grade 1 | 9 (7) | 2 (6) | |
| Grade 2 | 76 (58) | 23 (70) | |
| Grade 3 | 45 (35) | 8 (24) | |
| **Lymph Node Stage** ($n$ (%)) | | | 0.812 |
| Stage 1 | 0 (0) | 0 (0) | |
| Stage 2 | 111 (85) | 27 (82) | |
| Stage 3 | 19 (15) | 6 (18) | |
| **Tumour Size** *mm* | | | 0.362 |
| Median (Interquartile range) | 24 (16.25, 34) | 23 (15, 27) | |
| **Nottingham Prognostic Index** | | | 0.430 |
| Median (Interquartile range) | 4.86 (4.36, 5.50) | 4.46 (4.30, 5.50) | |
| **Prognositc Group** ($n$ (%)) | | | 0.997 |
| Good | 3 (2) | 1 (3) | |
| Moderate-1 | 35 (27) | 9 (27) | |
| Moderate-2 | 45 (35) | 11 (33) | |
| Poor | 37 (28) | 9 (27) | |
| Very Poor | 10 (8) | 3 (9) | |
| **Her2 Status** ($n$ (%)) | | | 0.546 |
| Unknown<br>Negative<br>Borderline<br>Positive | 15 (12)<br>98 (75)<br>4 (3)<br>12 (9) | 2 (6)<br>27 (82)<br>0 (0)<br>4 (12) | |

$0.606 \pm 0.026$ was achieved for grade prediction and a test accuracy of $0.824 \pm 0.012$ and test F1-score of $0.789 \pm 0.031$ was achieved for lymph node staging. For the direct approaches, fusion by self-attention achieved the highest weighted one-versus-rest mean test AUC of $0.786 \pm 0.020$. Fusion by average-pooling and fusion by Addition achieved the joint highest mean accuracy with $0.613 \pm 0.061$ and $0.613 \pm 0.080$ respectively, with fusion by average-pooling achieving a slightly lower standard deviation. Fusion by Addition got the highest F1-score of $0.611 \pm 0.078$ while fusion by average-pooling and self-attention achieving the next best F1-scores of $0.586 \pm 0.077$ and $0.584 \pm 0.066$ respectively. Direct prediction using just the primary tumour and no lymph node image achieved an AUC of $0.719 \pm 0.024$, accuracy of $0.544 \pm 0.038$ and F1-score of $0.532 \pm 0.038$. Although these results are worse than those previously mentioned, they are all within one standard deviation of both ABMIL with Concatenation and Hadamard product. Fig 5.5 shows the confusion matrix for each direct prediction model on the test set. For medical prediction tasks it is important for the recall of poor prognostic cases to be high so that none are missed. Additionally, precision of better prognostic cases should be high to avoid false negatives. Fusion by average-pooling achieved the highest recall of 0.80 for poor and very poor cases which is crucial as this is the most important class to identify. The approach also achieved the highest moderate-1 precision of 0.58 thus producing the fewest false negatives for that class. Across all methods poor prognostic cases had the highest recall of 0.689, moderate-1 had an average recall of 0.644 and moderate-2 had the lowest recall of 0.431. This suggests poor prognostic cases were the easiest to identify while moderate-2 cases were the hardest overall. This is not surprising as moderate-2 cases have similarities with both moderate-1 and poor prognostic groups making them harder to distinguish. Across all approaches moderate-2 cases were more likely to be predicted as moderate-1 instead of poor/ very poor.

Performance on the test set for prediction of moderate-1 vs moderate-2 vs poor/very poor using patches extracted at 20× can be seen in Table 5.4. Again, the part prediction approach achieved the highest accuracy of $0.703 \pm 0.034$ and weighted F1-score of $0.709 \pm 0.036$ compared to the direct prediction approaches. For the direct prediction approaches, fusion by self-attention achieved the highest accuracy of $0.581 \pm 0.064$ and weighted F1-score of $0.554 \pm 0.071$.

The model performances on the test set for prediction of moderate-1/moderate-2 vs poor/very poor using patches extracted at 10X can be seen in Table 5.3. Again the part prediction

Table 5.2: Test set mean AUC, Accuracy, and weighted F1-Score across 5-folds ± standard deviation moderate-1 vs moderate-2 vs poor/very poor using 10X magnification patches

| Method | AUC | Accuracy | F1-Score |
|---|---|---|---|
| Part Prediction | - | **0.727 ± 0.064** | **0.730 ± 0.064** |
| Direct(Primary Site Only) | 0.719 ± 0.024 | 0.544 ± 0.038 | 0.532 ± 0.038 |
| Direct+Concat | 0.726 ± 0.030 | 0.588 ± 0.100 | 0.567 ± 0.110 |
| Direct+Avgpool | 0.766 ± 0.010 | **0.613 ± 0.061** | 0.586 ± 0.077 |
| Direct+Hadamard | 0.722 ± 0.031 | 0.563 ± 0.052 | 0.534 ± 0.052 |
| Direct+Addition | 0.754 ± 0.020 | 0.613 ± 0.080 | **0.611 ± 0.078** |
| Direct+Self-attention | **0.786 ± 0.020** | 0.606± 0.047 | 0.584 ± 0.066 |

approach achieved the highest test set accuracy of 0.788 ± 0.069 and weighted F1-score of 0.789 ± 0.069 however this was only slightly higher than the direct prediction approaches. Fusion by self-attention achieved the highest performance across all metrics, achieving an AUC of 0.864 ± 0.015, accuracy of 0.769 ± 0.025 and weighted F1-score of 0.760 ± 0.028. ABMIL with Addition performed the second best with an AUC of 0.856 ± 0.030, accuracy of 0.756 ± 0.054 and weighted F1-score of 0.739 ± 0.079. Again direct prediciton using only a primary tumour site still achieved comparable results to all superior methods. Additionally, the approach outperformed fusion with Hadamard product across all metrics and achieved a higher accuracy and F1-score than average-pooling. Fig 5.6 shows the confusion matrix for each direct prediction model on the test set. Fusion by self-attention had the highest recall for the combined poor class at 0.65 and fusion by Hadamard product had the lowest with a poor/verypoor recall of 0.267. Fusion by self-attention had the highest precision for moderate cases with a precision of 0.8. Fusion by self-attention was the best approach for the binary classification task.

For prediction of moderate-1/moderate-2 vs poor/very poor using patches extracted at 20× patches again the part prediction approach was the best. It achieved an accuracy of 0.824 ± 0.035 and weighted F1-score of 0.822 ± 0.033. For the direct prediction methods, fusion by concatenation had the highest F1-score of 0.752 ± 0.045. Fusion by addition achieved the best accuracy of 0.781 ± 0.084 while the highest AUC was achieved by self-attention fusion, with a value of 0.876 ± 0.010.

Table 5.3: Test set mean AUC, Accuracy, and weighted F1-Score across 5-folds ± standard deviation moderate-1/moderate-2 vs poor/very poor using 10X magnification patches

| Method | AUC | Accuracy | F1-Score |
|---|---|---|---|
| Part Prediction | - | **0.788 ± 0.069** | **0.789 ± 0.069** |
| Direct(Primary Site Only) | 0.820 ± 0.011 | 0.725 ± 0.012 | 0.717 ± 0.007 |
| Direct+Concat | 0.829 ± 0.023 | 0.725 ± 0.013 | 0.720 ± 0.009 |
| Direct+Avgpool | 0.833 ± 0.053 | 0.706 ± 0.042 | 0.661 ± 0.093 |
| Direct+Hadamard | 0.760 ± 0.062 | 0.675 ± 0.064 | 0.585 ± 0.130 |
| Direct+Addition | 0.856 ± 0.030 | 0.756 ± 0.054 | 0.739 ± 0.079 |
| Direct+Self-attention | **0.864 ± 0.015** | **0.769 ± 0.025** | **0.760 ± 0.028** |

Table 5.4: Mean Accuracy, and weighted F1-Score across 5-folds ± standard deviation moderate-1 vs moderate-2 vs poor/very poor using 20X magnification patches

| Method | AUC | Accuracy | F1-Score |
|---|---|---|---|
| Part Prediction | - | **0.703 ± 0.034** | **0.709 ± 0.036** |
| Direct(Primary Site Only) | 0.764 ± 0.022 | 0.519 ± 0.032 | 0.498 ± 0.029 |
| Direct+Concat | 0.741 ± 0.020 | 0.550 ± 0.064 | 0.529 ± 0.111 |
| Direct+Avgpool | **0.779 ± 0.027** | 0.569 ± 0.072 | 0.493 ± 0.104 |
| Direct+Hadamard | 0.709 ± 0.042 | 0.519 ± 0.064 | 0.452 ± 0.074 |
| Direct+Addition | 0.739 ± 0.027 | 0.575 ± 0.015 | 0.554 ± 0.093 |
| Direct+Self-attention | 0.772 ± 0.041 | **0.581 ± 0.064** | **0.554 ± 0.071** |

Table 5.5: Mean Accuracy, and weighted F1-Score across 5-folds ± standard deviation moderate-1/moderate-2 vs poor/very poor using 20X magnification patches

| Method | AUC | Accuracy | F1-Score |
|---|---|---|---|
| Part Prediction | - | **0.824 ± 0.035** | **0.822 ± 0.033** |
| Direct(Primary Site Only) | 0.868 ± 0.024 | 0.744 ± 0.064 | 0.705 ± 0.117 |
| Direct+Concat | 0.850 ± 0.021 | 0.763 ± 0.042 | **0.752 ± 0.045** |
| Direct+Avgpool | 0.861 ± 0.059 | 0.706 ± 0.047 | 0.665 ± 0.094 |
| Direct+Hadamard | 0.752 ± 0.103 | 0.688 ± 0.97 | 0.585 ± 0.154 |
| Direct+Addition | 0.868 ± 0.017 | **0.781 ± 0.084** | 0.752 ± 0.139 |
| Direct+Self-attention | **0.876 ± 0.010** | 0.744 ± 0.013 | 0.729 ± 0.020 |

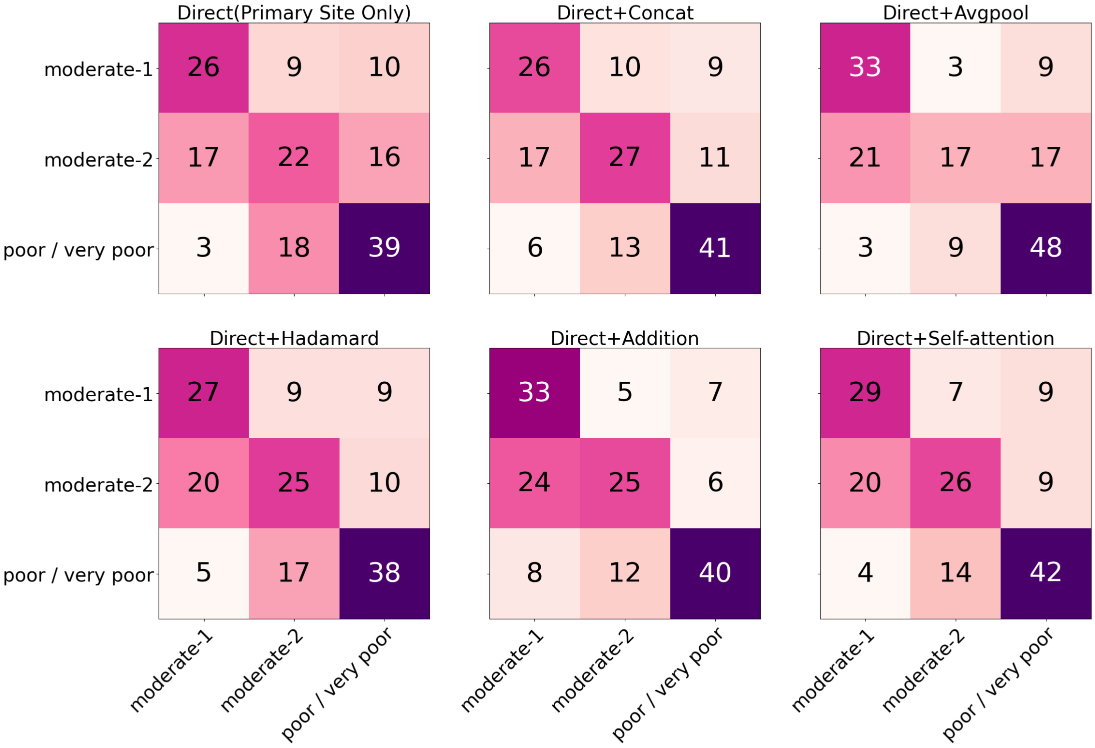


Figure 5.5: Confusion Matrices for all models trained for prediction between moderate-1, moderate-2 and poor/very poor classes using 10× patches. Left axis are the true labels and bottom axis are the predicted labels.

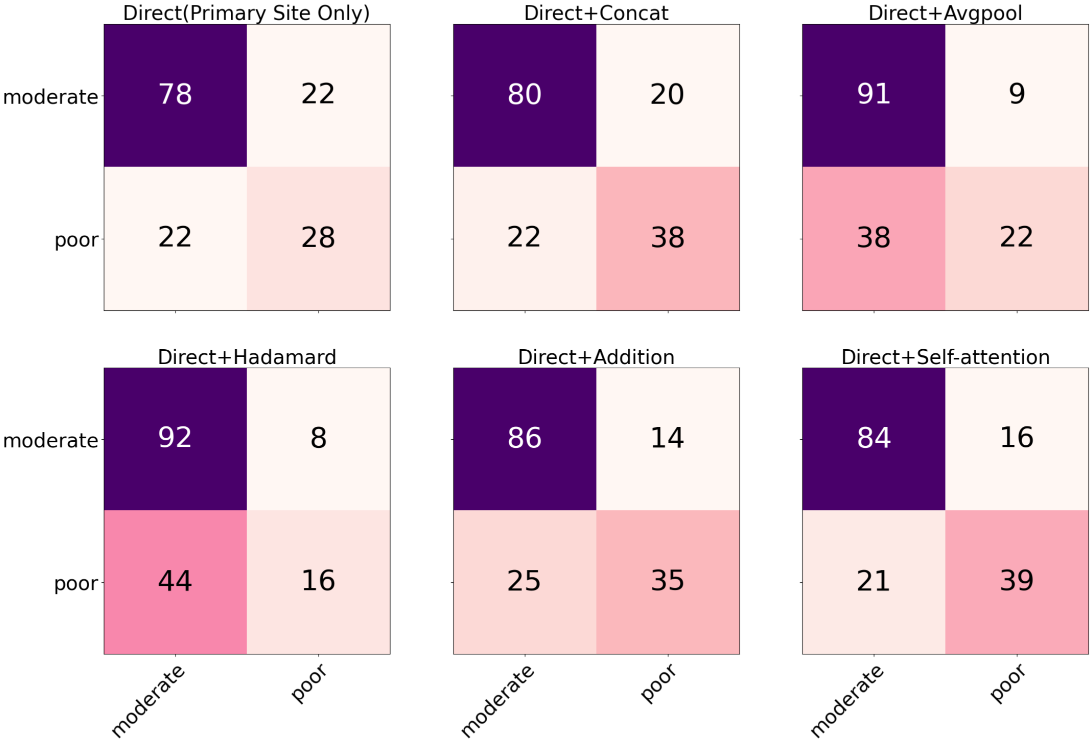


Figure 5.6: Confusion Matrices for all models trained for prediction between moderate-1/moderate-2 and poor/very poor classes using 10× patches. Left axis are the true labels and bottom axis are the predicted labels.

### Effect of Magnification on Model Performance

For the part prediction approach, using 10× magnification patches achieved a higher weighted F1-score compared to using 20× patches for prediction into three groups. On the other hand, using 20× patches with the part prediction approach achieved a higher F1-score for prediction into two groups. All direct prediction approaches performed best using 10× patches for prediction into three groups, while fusion with hadamard product and average pooling both benefited slightly from 20× for two group prediction.

The results show that using 10× magnification patches produces better results on average compared to using 20×. This suggests that patches extracted at 10× magnification provide the best balance between context and detail. The extra detail at the cost of context provided by higher magnification patches, seems to not be beneficial for prediction of the NPI group. Additionally, using lower magnifications results in fewer patches extracted per slide which significantly speeds up training and inference times. Overall, the results show that using 10× magnification patches is beneficial for NPI group prediction.

### Choice of Loss Function

Cross-Entropy Loss was used for prediction of NPI group, grading and lymph node staging. However, it could be hypothesised that an ordinal loss function would be better suited to the task. Using an ordinal loss function, the model prediction is still categorical however classes are ranked in order and the loss function is weighted by the difference in rank. Using breast cancer grading as an example, grade 1 is the lowest rank while grade 3 is the highest rank. If an input sample with ground-truth label grade 3 is predicted as grade 2 this would produce a smaller loss compared to if a prediction of grade 1 was given. This is because grade 3 is more similar to a grade 2 case than a grade 1 case. Using cross-entropy the same loss is incurred for an incorrect prediction of grade 2 and grade 1. The use of an ordinal loss was experimented with, specifically the method described in [35]. In the proposed approach the output of the model is no longer passed through a softmax but instead each individual output is passed through a sigmoid layer. This means each output node gives a value between 0 and 1. Additionally, the labels for data samples are changed to the form $(1, 1, ..., 1, 0, 0)$ in which the elements up to class $k$ are ones and the rest are zeros. The loss is then given by the mean squared error between the output neurons $O_i$ and the targets $t_i$. The classifier effectively must predict for the current class and all classes before it in the ranking. Therefore, for a target of $(1, 1, 1)$ a greater loss is produced for an output of $(1, 0, 0)$ compared to $(1, 1, 0)$.

However, from the experiments this form of ordinal loss function did not outperform the standard cross-entropy approach. Using the direct approach with self-attention fusion for moderate-1 vs moderate-2 vs poor/very poor group prediction with the ordinal loss on the validation sets achieved: AUC of $0.638 \pm 0.096$, accuracy of $0.529 \pm 0.094$ and weighted F1-score of $0.510 \pm 0.112$. In comparison, with the cross-entropy loss, the performance was: AUC of $0.655 \pm 0.103$, accuracy of $0.544 \pm 0.060$ and weighted F1-score of $0.527 \pm 0.075$. Similarly, using ABMIL with the primary site alone, the validation performance with ordinal loss was: AUC of $0.607 \pm 0.087$, accuracy of $0.534 \pm 0.119$ and weighted F1-score of $0.499 \pm 0.172$. With cross-entropy loss performance was: AUC of $0.681 \pm 0.085$, accuracy of $0.598 \pm 0.062$ and weighted F1-score of $0.594 \pm 0.060$.

### Number of Attention Blocks

The number of attention blocks used in the direct prediction approach did affect model performance. All results presented used a single attention block which took as input both the primary tumour patches and the lymph node patches. However, the use of 2 attention blocks was also explored. In this setup the first attention block would only process the primary site patches while the second block would process the lymph node patches. Results for the 2 attention block setup versus a single attention block can be seen in Table 5.6. Although using 2 attention blocks produced a slightly higher weighted F1-score on the validation set, in all cases both approaches were within 1 standard deviation of each other. Therefore, the single attention block approach was chosen due to the simpler model architecture resulting in fewer parameters and lower memory requirements.

Table 5.6: Validation set mean weighted F1-Score across 5-folds ± standard deviation moderate-1 vs moderate-2 vs poor/very poor using 10X magnification patches

| Model | 1 Attention Block | 2 Attention Block |
|---|---|---|
| Direct+Concat | 0.571 ± 0.084 | 0.567 ± 0.084 |
| Direct+Avgpool | 0.541 ± 0.103 | 0.553 ± 0.072 |
| Direct+Hadamard | 0.510 ± 0.086 | 0.569 ± 0.082 |
| Direct+Addition | 0.554 ± 0.102 | 0.579 ± 0.070 |
| Direct+Self-attention | 0.527 ± 0.075 | 0.529 ± 0.075 |

### 5.5.2 Interpretability and Attention Heatmap Visualisation

To visualise and understand the relative importance of each patch in the WSI to the final classification, attention heatmaps can be generated by converting the attention scores from the attention block for the predicted class of the model into percentiles. These normalised scores can then be mapped to their corresponding spatial location in the original slide. Fig 5.7 shows an attention map for a primary tumour and lymph node pair from each class. For each class the primary tumour image is shown on the top and the lymph node image is shown below. For each image three patches with the highest attention and three patches with the lowest attention are displayed alongside. The attention approach is capable of detecting the tumour tissue regions within both the primary site and lymph node. This is done without the need for ground-truth annotations and only using the NPI group label. Additionally, even though the same attention block processes both the primary tumour and lymph node images, the approach is still able to identify malignant tissue in both image types, despite the difference in appearance. In the moderate-1 case high attention is given to regions of dense, irregular nuclei within the primary tumour while low attention is given to fat stromal regions. In the corresponding lymph node the high attention patches highlight regions or cancerous nuclei with large quantities of tumour infiltrating lymphocytes (TILS), indicating the immune response to the metastasis. In the moderate-2 cases a comparatively smaller tumour is found within the centre of the primary tumour slide. High attention is given to this region despite the majority of the slide containing stromal tissue and several healthy glands. Attention is also given to smaller regions of invasive, as well as in-situ carcinoma above the largest tumour region. Within the lymph node high attention patches contain irregular epithelial cells indicating metastatic cancer. Similar to the moderate-1 case, TILS are present within these patches, although not as prevalent. In the poor case regions of irregular infiltration of tumour cells within the breast tissue are given the highest attention. There is no formation of tubule structures and counts of mitotic figures are high indicating a higher grade cancer. Within the lymph node, regions or large epithelial cells with nuclear irregularity are given high attention within the lymph node. The high attention regions also contain high counts of mitotic figures and very few TILS which suggests the presence of an aggressive cancer. Similar to all cases the lowest attention is given to fat, stromal regions which contain limited prognostic information.

#### Does the direct approach encode for Grade and Lymph Node stage?

The direct prediction approach aimed to predict the NPI group directly from the slides without additional information from the grade and lymph node stage labels. However, determining whether these factors can be discovered from the learned model features could provide greater interpretability and utility to the model explanations. To achieve this, principles from concept attribution methods (section 2.6.4) were used.

To determine if a model layer has learned the concepts of grade and lymph node stage, linear probing can be used. For this, a linear model can be fitted to the bag embeddings to determine if the concepts of grade or lymph node stage have been learned. The degree to which a linear model can predict the presence of certain concepts indicates how well the model can recognise them. A Support Vector Machine (SVM) with hinge loss and $L1$ regularisation was used for this. The SVM was fitted to predict either grade or lymph node stage using the bag embeddings from either the primary tumour site, lymph node or their combined output after concatenation, average-pooling, addition, hadamard product or self-attention head. The model weights were frozen during this, and only the SVM was trained.

For prediction of the grade concept, only fusion by self-attention achieved an accuracy and weighted F1-score greater than 0.5 using the primary site embedding. This suggests that this concept was not recognised. No methods using the combined embedding achieved a high enough accuracy or F1-score to suggest the concept of grade was learned. Surprisingly, fusion by ad-

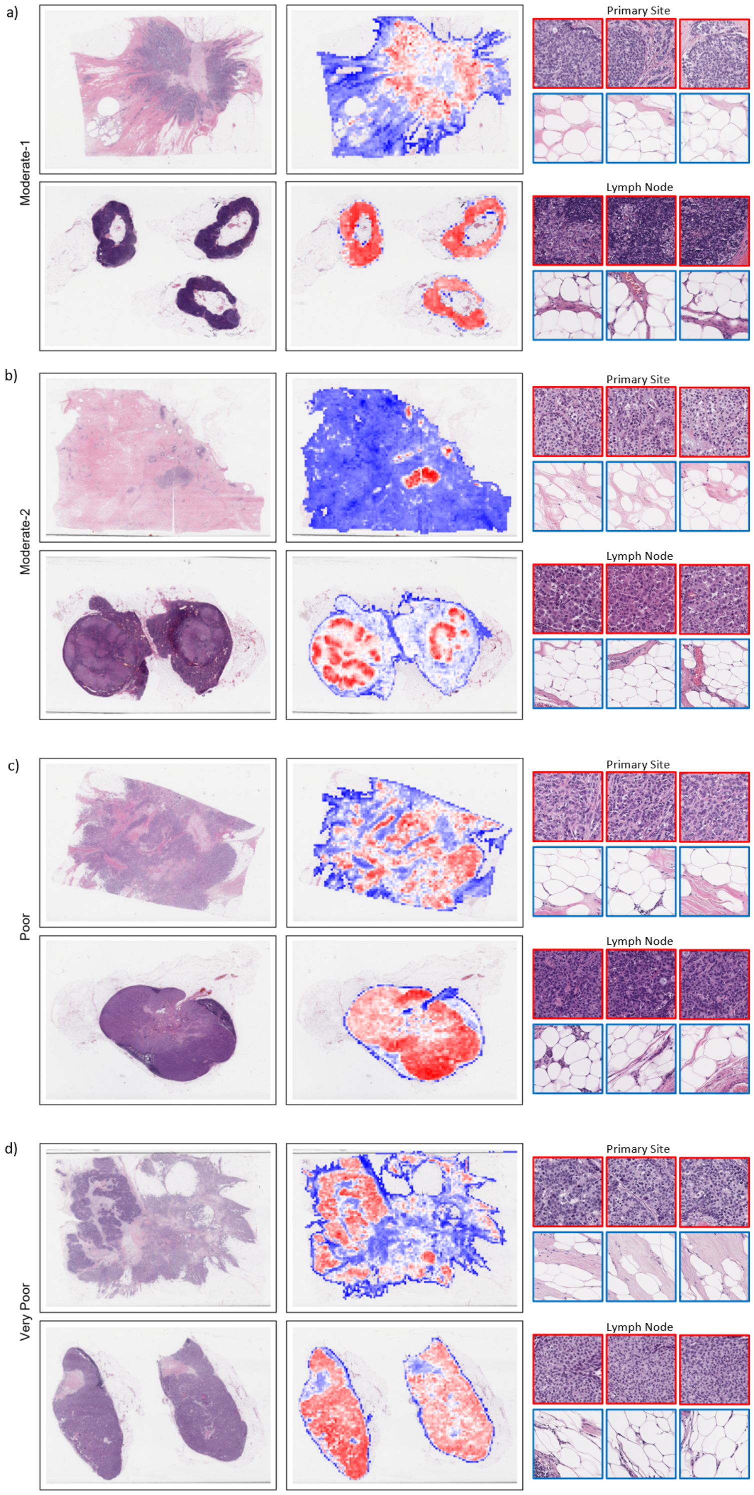


Figure 5.7: Visualisation of attention heatmaps for a) moderate-1, b) moderate-2, c) poor and d) very poor prognosis cases. Attention maps are taken from the ABMIL with self-attention model trained for classifying between moderate-1/moderate-2 and poor/very poor cases. Left column shows primary tumour site and lymph node pairs. Middle column shows overlaid attention heatmaps. Blue indicated low attention, white indicated moderate attention and red indicated higher attention. Right column shows the three patches with the highest attention (red border) and three patches with the lowest attention (blue border) for the primary tumour and lymph node.

dition achieved an accuracy and weighted F1-score of 0.582 and 0.563 respectively for grade prediction using the lymph node embedding. For lymph node stage prediction all approaches were able to distinguish between lymph node stages when using the lymph node embeddings. Fusion by self-attention achieved the highest accuracy and F1-score of 0.830 and 0.834 respectively when using the lymph node embedding. A summary of these results can be seen in Fig 5.7.

The greater performance on the lymph node staging task over breast cancer grading could be explained in two ways. First, the task of lymph node staging is likely an easier task compared to breast cancer grading. This is because the staging was treated as a binary classification problem as only two different stages were present in the dataset. On the other hand, breast cancer grading was a three class problem making it slightly more different especially when grade 1 was very under represented in the dataset. Second, the results could suggest that the lymph node bag embeddings are more informative than the primary site embeddings, which would give greater performance for lymph node staging. This might also explain why using fusion by addition with the lymph node embeddings achieved the best grading performance. This could further suggest that the lymph node bag embeddings were more important than the primary site embedding for NPI group prediction as the lymph node embeddings features are more informative.

Table 5.7: Mean Accuracy, and weighted F1-Score across 5-folds ± standard deviation for prediction of breast cancer grade and lymph node stage from the embeddings learned through direct prediction of NPI group.

| | Method | Primary | | Lymph | | Combined | |
|---|---|---|---|---|---|---|---|
| | | Acc | F1 | Acc | F1 | Acc | F1 |
| Grade | Direct(Primary Site Only) | 0.267 ± 0.225 | 0.200 ± 0.219 | - | - | - | - |
| | Direct+Concat | 0.212 ± 0.124 | 0.150 ± 0.130 | 0.478 ± 0.172 | 0.472 ± 0.197 | 0.273 ± 0.131 | 0.246 ± 0.139 |
| | Direct+Avgpool | 0.248 ± 0.106 | 0.210 ± 0.123 | 0.424 ± 0.133 | 0.429 ± 0.144 | 0.448 ± 0.119 | 0.434 ± 0.172 |
| | Direct+Hadamard | 0.491 ± 0.205 | 0.482 ± 0.210 | 0.406 ± 0.176 | 0.365 ± 0.157 | 0.333 ± 0.103 | 0.278 ± 0.148 |
| | Direct+Addition | 0.394 ± 0.145 | 0.366 ± 0.160 | 0.582 ± 0.181 | 0.563 ± 0.188 | 0.406 ± 0.131 | 0.375 ± 0.174 |
| | Direct+Self-attention | 0.576 ± 0.057 | 0.577 ± 0.055 | 0.273 ± 0.237 | 0.254 ± 0.248 | 0.473 ± 0.156 | 0.442 ± 0.204 |
| LN Stage | Direct(Primary Site Only) | - | - | - | - | - | - |
| | Direct+Concat | 0.709 ± 0.133 | 0.725 ± 0.125 | 0.745 ± 0.104 | 0.771 ± 0.093 | 0.764 ± 0.052 | 0.781 ± 0.035 |
| | Direct+Avgpool | 0.655 ± 0.223 | 0.630 ± 0.257 | 0.788 ± 0.140 | 0.784 ± 0.130 | 0.661 ± 0.098 | 0.694 ± 0.096 |
| | Direct+Hadamard | 0.648 ± 0.239 | 0.633 ± 0.291 | 0.691 ± 0.164 | 0.711 ± 0.174 | 0.527 ± 0.268 | 0.505 ± 0.309 |
| | Direct+Addition | 0.667 ± 0.204 | 0.648 ± 0.198 | 0.648 ± 0.158 | 0.676 ± 0.149 | 0.776 ± 0.128 | 0.785 ± 0.109 |
| | Direct+Self-attention | 0.485 ± 0.209 | 0.478 ± 0.241 | 0.830 ± 0.062 | 0.834 ± 0.055 | 0.836 ± 0.036 | 0.844 ± 0.029 |

## 5.6 Conclusion

To summarise, two approaches for prediction of the NPI group have been explored and compared. The first approach aimed to predict both grade and lymph node independently and combined the results with the pathologist reported invasive tumour size to predict the NPI group. The second approach looked at the potential of directly predicting the NPI group by combing the primary site and lymph node slide information through late fusion. A range of fusion operators were explored and the used of addition or a self-attention module proved to be the best performers.

The first approach achieved superior performance compared to the latter. Models trained for specific prediction of each component are more tailored to a certain task and can therefore outperform a more general approach. Additionally, the use of pathologist reported tumour size benefited performance as this was something the model did not need to account for. However, this does mean that the approach is not a completely automated prediction tool as the pathologist is still required to measure the tumour size. On the other hand, direct prediction of the NPI group does not require any information provided by the pathologist. Instead the approach is capable of automatically predicting NPI group without needing pathologist input. Additionally, such an approach used a single primary site WSI and single lymph node image while in practice a pathologist would need to look at significantly more slides to make a judgement. Therefore, such a method could reduce the number of slides required per patient. The approach does suffer from weaker performance however and is not capable of predicting patient grade. Prediction of lymph node stage was found to be possible by training a new classifier on the lymph node slide embeddings produced by the model.

This work has built upon other data fusion methods by exploring the use of combining a patient's primary site and lymph node information. By introducing data fusion, model performance was increased compared to using a single slide. Combining multiple slides from a single patient might have a real benefit and should be explored further for other tasks such as survival analysis. NPI group is an established indicator for patient 10-year survival suggesting that the benefits of combining slides shown here is relevant to the area of survival analysis.

Overall, the performance of neither approach is adequate for clinical deployment. The dataset size was a limiting factor with only 163 sample pairs. Additionally, some classes were underrepresented such as the good prognostic group or cases with a lymph node stage of 1 (no cancer in lymph nodes). In order to improve performance a larger dataset would be required with inclusion of underrepresented classes. Additionally, more sophisticated approaches to the multiple instance problem could be explored such as graph neural networks or transformers which account for spatial location and dependencies between patches.

# Chapter 6

# Cross-Attention Multiple Instance Learning

## 6.1 Introduction

Although attention-based approaches provide some insight into the model's decision making this is often restricted to a heatmap of important instances. Prototype-based MIL approaches have been proposed to provide greater interpretability [165, 145, 117, 58]. Similar to bag-of-words approaches, prototype-based MIL represents a WSI by comparing it against a set of prototypes. The aim is for each prototype to capture a different morphology within the slide. Here a morphology refers to both homogeneous regions, containing a single tissue type with cells of similar size and shape, as well as more complex structures, a region of both tumour and stroma, or regions of tumour infiltrating lymphocytes which would contain a variety of cellular types. The WSI can then be represented as the histogram of counts of different prototypes which can be used for classification. Alternatively, the features from different morphological slide regions can be aggregated together or their spatial arrangements can be considered.

Here a prototype-based MIL approach is proposed. A set of prototypes, or queries, are learned and then compared against patches from a WSI. This represents a slide as regions with different tissue and cell structures. An attention mechanism is subsequently used to determine the saliency of each morphological region towards the final prediction. This work differs from other prototype-based MIL approaches in several ways. First, while other methods pre-compute the prototypes, usually through unsupervised clustering (e.g $k$-means), here the set of prototypes are learned end-to-end. This allows for more refined prototype creation. Secondly, patches are often only assigned to the most similar prototype in the dictionary, however here patches are represented as a weighted assignment to each prototype. This is important, particularly at lower magnifications, when patches can contain multiple tissue and cell types and therefore cannot be represented by a single prototype. Finally, the use of a cross-attention layer provides the benefit of a learnable similarity measure between patches and prototypes, as apposed to a fixed metric such as euclidean distance. The approach is validated on both lung cancer sub-type prediction and breast cancer grading and competitive performance with other SOTA MIL approaches is found.

The main contributions of this chapter are: (1) A prototype-based MIL approach based on cross-attention is proposed to provide greater interpretability to MIL methods. (2) Unlike other prototype-based MIL approaches, here the prototypes are learned end-to-end during training removing the requirement for a prototype discovery step. In addition, the patches from a WSI are represented as a weighted assignment to each prototype and a learnable similarity measure is introduced. (3) The Cross-Attention MIL achieves comparable performance to other SOTA MIL methods on lung cancer sub-type prediction and achieves superior performance in breast cancer grading.

## 6.2 Related Works

There is increasing interest in the role prototypes can play within MIL methods. Prototypes can help increase model interpretability by disentangling the WSI into core components, they can help refine the feature space to cluster around key morphological structures and also detect rare cases. To increase model interpretability, Vu et al [145] proposed the Handcrafted Histological Transformer (H2T) for unsupervised representation learning of WSI. They took inspiration from the design choices behind the popular transformer architecture and used prototypes for their unsupervised representations. The method first extracts a set of prototypical patterns using a set of reference WSI. In inference, WSI are projected against the learned prototypical patterns. The patterns across the slide can then be summarised for downstream tasks. Several methods are used to summarise the patterns such as a co-occurrence matrix or a CNN. Yu et al [165] introduced the Prototypical multiple instance approach specifically for prediction of lymph node metastasis with a focus on improving detection of micro-metastases, which are smaller tumour cell clusters with a diameter of $0.2mm$ to $2mm$. The method was inspired by the traditional Bag-of-visual-words (BoVW) [76] which uses the occurrence of counts of a vocabulary of local image features to encode an image. They use a prototype discovery module to learn a set a prototypes via clustering the latent features. They then represent a slide as a histogram of assignment maps for each patch against each prototype. They use a learnable similarity metric inspired by metric learning to determine the similarity between a prototype and a patch. The similarity assignment vector is then used for classification. This prototypical MIL outperformed other methods such as ABMIL and DS-MIL in micro-metastases detection.

## 6.3 Methods

### 6.3.1 Overview

Patches $x$ are extracted from tissue regions within the slide, which can then be represented as $X = \{x_1, ..., x_L\}$ where $L$ can vary in size. Following conventional MIL approaches, each patch is first processed using a ResNet-50 [60] pretrained on ImageNet. This reduces each image patch to a feature embedding $h \in \mathbb{R}^d$ where $d$ is the size of the embedding vector. For a given WSI, each patch feature vector $h_j$, where $j$ is the index of the patch, is passed through a FC layer to reduce it to a size $d'$ which is then compared against a set of $m$ prototypes $P = \{p_n\}_{n=1}^{m}$. This produces a set of similarity scores determined using the cross-attention layer. For each prototype, the similarity scores between it and all patches are passed through a softmax function. This returns $m$ heatmaps displaying the assignment of each prototype across the WSI. For each prototype, the similarity scores are used to compute a weighted average with each patch vector. This produces feature vectors $C = \{c_n\}_{n=1}^{m}$. These feature vectors therefore represent the slide information most similar to the prototype. Finally an attention mechanism is used to determine what slide information is most important for a particular classification. This works by, for each class $cls$, returning an attention coefficient $a_{cls,n}$ for each feature vector $c_n$. A weighted average of all feature vectors is produced to generate the final slide feature vector per class. The attention coefficients are calculated using multi-branch attention [98] and the final slide feature vector per class is given by

$$z_{cls} = \sum_{n=1}^{m} a_{cls,n} c_n \tag{6.1}$$

where:

$$a_{cls,n} = \frac{exp\{W_{cls}(tanh(Vc_n^\top) \odot sig(Uc_n^\top))\}}{\sum_{j=1}^{m} exp\{W_{cls}(tanh(Vc_j^\top) \odot sig(Uc_j^\top))\}.} \tag{6.2}$$

Here $V$, $U$ are weight matrices shared across all classes however $W_{cls}$ are parallel branches equal to the number of classes. This allows for identification of the important slide features per class. This results in a slide feture vector per class which are passed through independent classifier layers resulting in a score for each class. The full architecture can be seen in Fig 6.1.

### 6.3.2 Cross-Attention Layer

To assign patches to different prototypes a cross-attention layer is used, inspired by the self-attention method in transformers. Here a set of prototypes are queried against the slide to locate the expression of different morphologies. To do this, first the patch embeddings are mapped to a key $K$ and value $V$ vector using linear layers $W_k \in \mathbb{R}^{d \times d'}$ and $W_v \in \mathbb{R}^{d \times d'}$. Then a set of prototypes are initialised to $m$ random vectors of length $d'$ before training. The output of the slide when queried against a particular prototype $p_n$ is therefore

$$c_n = softmax(\frac{p_n K^\top}{\sqrt{d'}})V \tag{6.3}$$

where K and V are the key and values vectors for each patch and $\sqrt{d'}$ is a scaling factor. The output of the cross-attention layer can be thought of as a vector which summaries the slide information most similar to a particular prototype. By visualising the similarity scores for each prototype the slide can be segmented into distinct regions containing distinct cell and tissue structures.

### 6.3.3 Prototype Learning

To learn a range of prototypes which resemble distinct morphologies two terms are added to the loss function when training. Alongside cross-entropy for the classification loss, a cluster loss [52] and orthogonality loss [146] are included. Both loss terms help create diverse and relevant prototypes which activate across different slide regions. The cluster loss encourages each prototype to be the most similar to at least one patch within a WSI. To achieve this, for each patch, the similarity score between it and each prototype is softmaxed returning $m$ similarity maps $s_m \in S$ where $S = softmax(\frac{PK^\top}{\sqrt{d'}})$. Then the maximum similarity score for each prototype across the slide is taken. This gives $m$ maximum scores which are then averaged to return the loss given by,

$$\mathbb{L}_{clst} = -\frac{1}{m}\sum_{n=1}^{m} max_{s_n \in S}(s_n) \tag{6.4}$$

The orthogonal loss encourages each prototype to be orthogonal to one another and thus promotes diversity between prototypes. The orthogonal loss forces the matrix multiplication of each pair of prototype vectors to be as close as possible to the identity matrix. This is given by,

$$\mathbb{L}_{orth} = ||PP^\top - I||^2 \tag{6.5}$$

where $I$ is the identity matrix and $P$ are the prototypes which are normalized to unit length. The final loss function is therefore,

$$\mathbb{L} = \mathbb{L}_{CE} + \lambda_{clst}\mathbb{L}_{clst} + \lambda_{orth}\mathbb{L}_{orth} \tag{6.6}$$

## 6.4 Experiments

The Cross-Attention MIL is validated on two histopathology WSI classification tasks and the results are compared against several other SOTA methods [98, 67, 122]. Additionally, the end-to-end prototype learning is compared against k-means clustering on the training set.

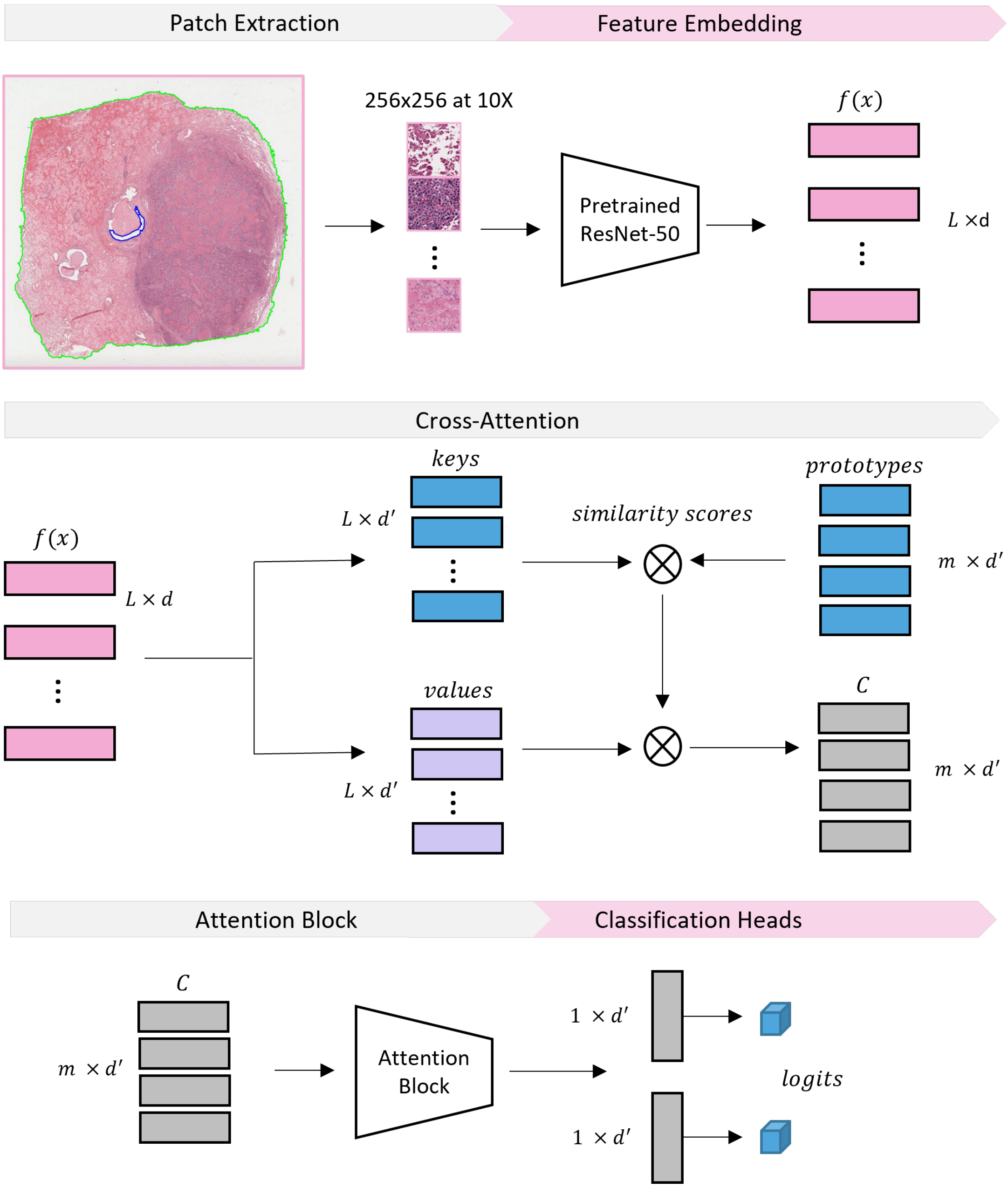


Figure 6.1: Illustration of the proposed Cross-Attention MIL. Patches $x$ are extracted and passed through a ResNet50 $f$ pretrained on ImageNet. The patch embeddings $h$ are then mapped to a set of *keys* and *values*. The *keys* are compared against a set of $m$ *prototypes* and the *values* are multiplied by the resulting similarity scores. This produces $m$ vectors $C$ which are passed through an attention block to determine the features associated with each class.
.

### 6.4.1 Datasets

To validate the proposed Cross-Attention MIL approach two classification tasks as performed. The first task is lung cancer sub-type classification (https://portal.gdc.cancer.gov/). For this data from TCGA open-access portal was used. This dataset is summarised in chapter 3. The second task is prediction of breast cancer grade between low/moderate (grade 1 & grade 2) and high grade (grade 3) cases. For this we use 163 WSI obtained from LTHT. The dataset contains 11 low grade slides, 99 moderate grade slides and 53 high grade slides. Due to the limited number of low grade slides they were combined with the moderate grade class. This effectively made the classification problem low/moderate grade vs high grade. For both datasets non-overlapping patches of size 256 × 256 pixels were extracted from all non-background slide regions. Patches are extracted at both 10× and 20× magnification to evaluate the proposed approach at different magnifications.

### 6.4.2 Implementation Details

In each dataset 80% of samples were used for training and validation while the remaining 20% were used for testing. 5-fold cross validation was also implemented. Adam optimiser was used with a fixed learning rate of $1 \times 10^{-4}$ and a weight decay of $1 \times 10^{-5}$. Cross Entropy Loss was used for the slide label loss. The models were trained for a maximum of 30 epochs. To determine the optimal hyperparameters, different numbers of prototypes were swept over {6, 8, 10, 12}, as well as several coefficients for the cluster cost {0.2, 0.4, 0.6, 0.8, 1.0} and orthogonality cost {0.2, 0.4, 0.6, 0.8, 1.0}. The hyperparameters with the best performance on the first fold of the validation set were selected for testing and can be seen in Table 6.1.

For comparison with ABMIL and CLAM, the same optimiser, learning rate and weight decay were selected. Additionally, the bag weight in CLAM was set to 0.7. For training of TransMIL the Lookaheadoptimizer [172] was deployed with a learning rate of $2 \times 10^{-4}$ and weight decay of $1 \times 10^{-5}$ as recommended in their work. Training was implemented with a single NVIDIA V100 GPU on the Advanced Research Computing 4 (ARC4) system at the University of Leeds. The models were implemented in PyTorch [108].

Table 6.1: Selected hyperparameters for Cross-Attention MIL

| Experiment | Prototypes | Cluster ($\lambda_{clst}$) | Orthogonality ($\lambda_{orth}$) |
|---|---|---|---|
| Lung 10× | 8 | 0.8 | 1.0 |
| Lung 20× | 12 | 0.2 | 1.0 |
| Breast 10× | 8 | 0.6 | 1.0 |
| Breast 20× | 12 | 1.0 | 1.0 |

## 6.5 Results and Discussion

### 6.5.1 Cross-validated Model Performance

The performance of Cross-Attention MIL compared with other methods for lung cancer subtype prediction, can be seen in Table 6.2. The results show that the Cross-Attention approach slightly outperforms the other benchmark methods with $\approx$ 1% increase in accuracy and F1-score over TransMIL and CLAM and a much smaller increase over ABMIL. Using patches at 20× magnification, ABMIL was the best performing method achieving the highest accuracy and weighted F1-score. Cross-Attention MIL did achieve the joint highest AUC of 0.929 along with TransMIL. Cross-Attention MIL was still within two standard deviations of ABMIL and is therefore still competitive with other SOTA methods.

For breast cancer grading (Table 6.3) Cross-Attention MIL at 10× achieves a significantly higher performance compared to the other methods. Cross-Attention MIL achieved an increased accuracy of $\approx$ 6% compared to CLAM, $\approx$ 1% compared to ABMIL and $\approx$ 7% compared to TransMIL. The weighted F1-score also increases by $\approx$ 5% over CLAM, $\approx$ 3% over ABMIL and $\approx$ 6% over TransMIL. However, at 20× magnification CLAM outperformed Cross-Attention with a $\approx$ 1% increase in weighted F1-score. Cross-Attention MIL was still the second best performing method for breast grading using 20× patches with a $\approx$ 3% increase in weighted F1-score over ABMIL and $\approx$ 1% over TransMIL. From both datasets it is clear that Cross-Attention MIL achieves better performance when using lower magnification patches. All other methods also perform better with 10× magnification patches. This agrees with the findings in section 5.5.1 and further suggests that 10× magnification patches provide the best balance

Table 6.2: Test set mean AUC, Accuracy, and weighted F1-Score across 5-folds ± standard deviation for LUAD vs LUSC classification

| Magnification | Method | AUC | Accuracy | F1-Score |
|---|---|---|---|---|
| 10× | ABMIL | 0.920 ± 0.008 | 0.873 ± 0.020 | 0.873 ± 0.020 |
| | CLAM | 0.917 ± 0.010 | 0.867 ± 0.008 | 0.867 ± 0.008 |
| | TransMIL | 0.925 ± 0.011 | 0.864 ± 0.018 | 0.863 ± 0.018 |
| | Cross-Attention | **0.936 ± 0.012** | **0.875 ± 0.019** | **0.875 ± 0.019** |
| 20× | ABMIL | 0.928 ± 0.009 | **0.869 ± 0.011** | **0.869 ± 0.012** |
| | CLAM | 0.927 ± 0.009 | 0.861 ± 0.008 | 0.861 ± 0.008 |
| | TransMIL | 0.929 ± 0.015 | 0.848 ± 0.014 | 0.848 ± 0.014 |
| | Cross-Attention | **0.929 ± 0.006** | 0.857 ± 0.006 | 0.857 ± 0.007 |

between context and detail.

The lower performance of TransMIL on the breast cancer grading dataset is surprising, particularly when using using 10× magnification patches. The lower performance could be due to the smaller dataset in comparison to the lung dataset. TransMIL might require more data to train as dependencies between patches must be learned compared to other methods. It is likely the case that, had there been more data for breast cancer grading, TransMIL would have performed better or similar to other approaches.

In general, the models all perform equivalently on the lung sub-typing task while performance was more varied for the breast grading task. This is likely for two reasons. The first is that lung sub-typing is an easier task as each sub-type displays a distinct morphological pattern. Breast cancer grading is instead a spectrum where some features can be shared across grades while others might differ. This might make grading a more difficult task with more variation. Secondly, the lung sub-typing dataset has more slides available for training and evaluation which means that all models will likely perform better across the board. Therefore, this could suggest that one benefit of Cross-Attention MIL is the ability to perform better on smaller datasets compared to other methods.

Table 6.3: Test set mean AUC, Accuracy, and weighted F1-Score across 5-folds ± standard deviation for low/moderate vs high breast cancer grade classification

| Magnification | Method | AUC | Accuracy | F1-Score |
|---|---|---|---|---|
| 10× | ABMIL | 0.720 ± 0.022 | 0.745 ± 0.031 | 0.731 ± 0.047 |
| | CLAM | 0.707 ± 0.030 | 0.700 ± 0.027 | 0.715 ± 0.026 |
| | TransMIL | **0.727 ± 0.033** | 0.685 ± 0.091 | 0.698 ± 0.082 |
| | Cross-Attention | 0.725 ± 0.015 | **0.758 ± 0.000** | **0.761 ± 0.009** |
| 20× | ABMIL | 0.692 ± 0.029 | 0.679 ± 0.041 | 0.641 ± 0.041 |
| | CLAM | 0.683 ± 0.029 | 0.679 ± 0.071 | **0.686 ± 0.055** |
| | TransMIL | **0.734 ± 0.019** | 0.648 ± 0.071 | 0.666 ± 0.063 |
| | Cross-Attention | 0.702 ± 0.027 | **0.703 ± 0.048** | 0.673 ± 0.051 |

### 6.5.2 Model Interpretability

To interpret a model decision, the attention heatmaps can be visualised to determine which slide regions were the most influential. Additionally, the prototype assignment maps can be shown as well. This map shows the closest prototype to each patch providing an understanding of the different tissue regions within a slide. To create the attention heatmaps, the final attention value of a patch is given by the sum of each prototype's attention value multiplied by the similarity between the patch and the prototype.

An example for prediction of both lung sub-types can be seen in Fig 6.2 along with images of the closest patch to each prototype from the image. Looking at the closest patches to each prototype, numbered from left to right with the left most patch being number 1 and the right most patch being number 8, prototypes 2 and 4 contain malignant tumour cells. Prototype 1 looks for regions of adipose. Prototype 5 contains regions of connective tissue and prototype 8 also shows regions of connective tissue but with the presence of red blood cells. The heatmaps appear to localise the tumour regions in both cases with prototype 2 receiving the most attention.

An example for prediction of both grade 2 and grade 3 breast cancer can be seen in Fig 6.3. From the figure it can be seen that prototype 2 contains high densities of tumour cells. Prototype 8 contains immune cells as well as some tumour cells showing an immune response to the cancer. Prototypes 1, 3 and 7 contain connective tissue with varying amounts of adipose.

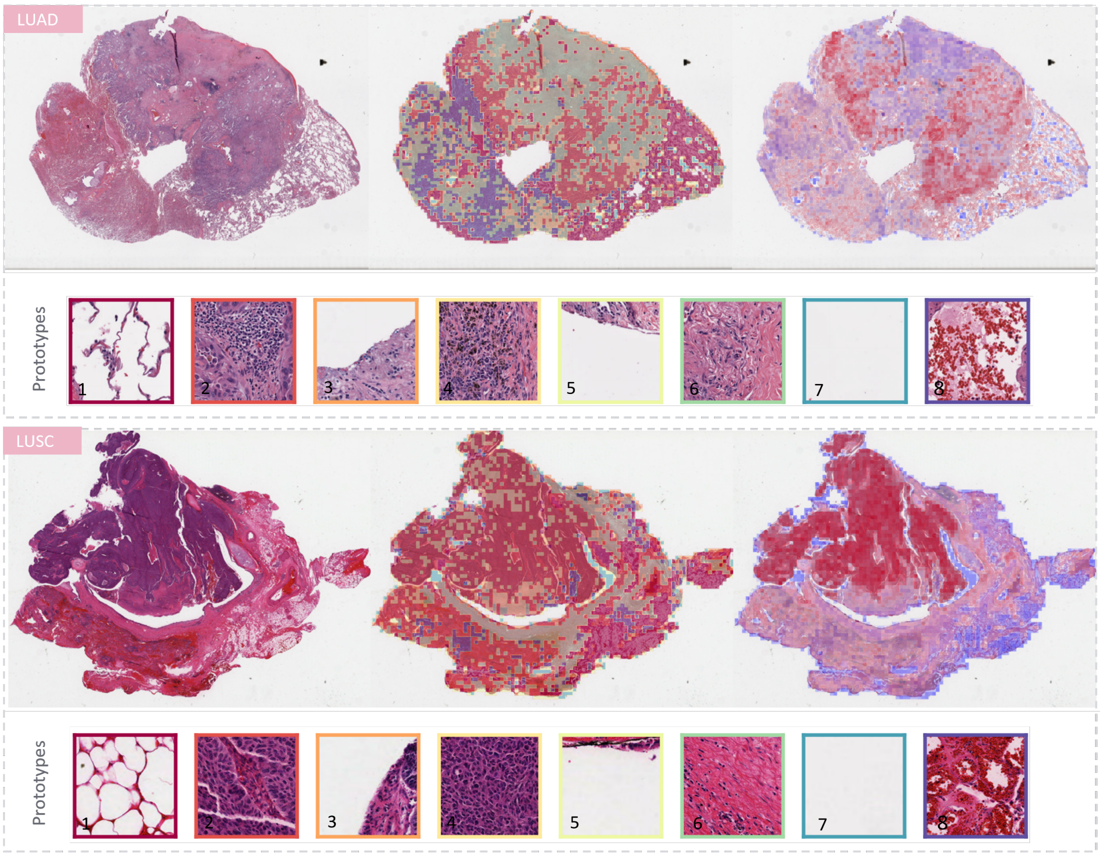


Figure 6.2: Cross-Attention MIL attention maps for LUAD and LUSC cases. Top: (Left) Original WSI, (Middle) prototype similarity map, (Right) attention heatmap with higher attention in red and lower attention in blue. Bottom: this shows an example patch closest to each prototype. Borders around each example patch refer to their location in the prototype similarity map.

.

Prototypes 4, 5 all contains dense connective tissue with varying amounts of epithelial, tumour and lymphatic cells. Prototype 6 is also stromal tissue but with very low density of cells. The attention heatmaps localise well the tumour regions within the slide thus giving prototype 2 the highest attention in both grade 2 and grade 3 cases.

In addition to the visualised attention heatmaps, the average attention given to each prototype representation $c_n$ across each class can be found. This provides a global explanation instead of a local one, allowing for an understanding of the important prototypes per class. For Lung cancer sub-type prediction using 10× patches, the average attention for each prototype can be seen in Fig 6.4. This figure shows the results on the test set using the model trained

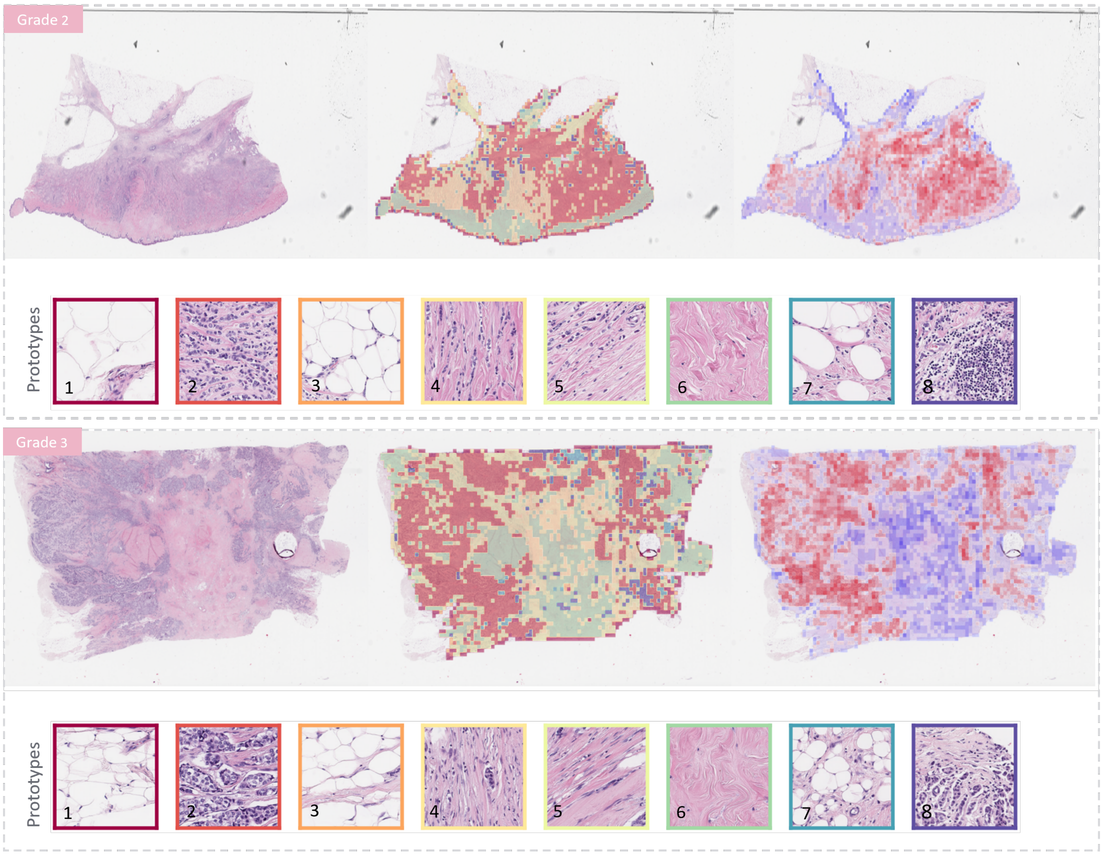


Figure 6.3: Cross-Attention MIL attention maps for a grade 2 and grade 3 cases. Top: (Left) Original WSI, (Middle) prototype similarity map, (Right) attention heatmap with higher attention in red and lower attention in blue. Bottom: this shows an example patch closest to each prototype. Borders around each example patch refer to their location in the prototype similarity map.

.

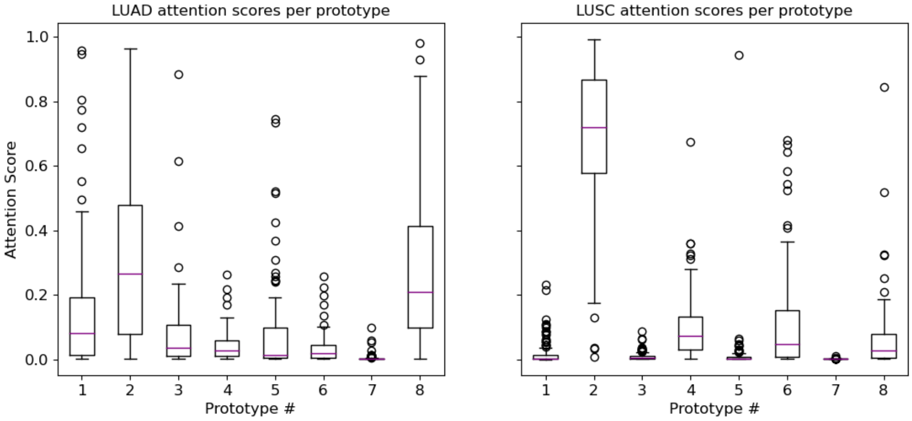


Figure 6.4: Box plots of attention score per prototype. Left is for LUAD predictions. Right is for LUSC predictions.
.

using the first fold. From the boxplots it can be seen that prototype 2 has a much higher median value for LUSC predictions compared to LUAD ones. The median of prototype 2 for each class also lies outside the interquartile range of the opposing boxplot suggesting a large difference between the two groups. Prototypes 1 and 8 show a higher median for LUAD predictions compared to LUSC predictions where the median is almost 0. For Breast cancer grading using 10× patches, the average attention for each prototype can be seen in Fig 6.5. Here prototype 2 is given a much higher attention for higher grade predictions compared with low and moderate grade predictions, however is still important for low grade predictions just to a lesser degree. All other prototypes are given a small amount of attention ($< 0.2$) for low and moderate grade predictions and almost 0 attention for high grade predictions.

### Prototype Similarity map features

As outlined in section 2.6.4 various methods have evaluated human-interpretable features for different diagnostic tasks. These methods often first require prediction of predetermined tissue and cell types. Following this, shape, boundary features or co-localisation of components can be extracted as features for prediction tasks. However, these approaches often require large amounts of labelled data for accurate segmentation and classification as well as knowledge of the features that will likely be of interest. Prototype similarity maps could be used to extract

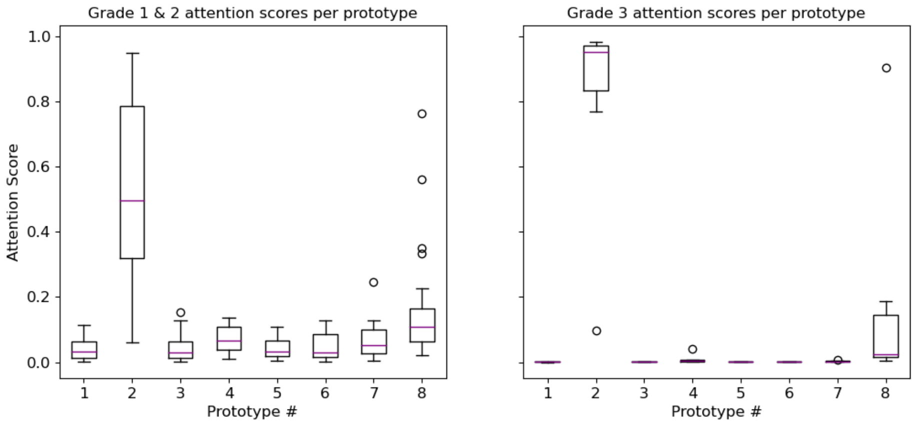


Figure 6.5: Box plots of attention score per prototype. Left is for Grade 1 & 2 predictions. Right is for Grade 3 predictions.
.

similar metrics, such as co-localisation features or amounts of certain slide morphologies, but without the need for large sets of annotated data or initial domain knowledge. Therefore, in this section the prototype similarity maps are assessed for their ability to extract human-interpretable features.

To do this, several slide level features were measured and the difference between the features per class were found. To extract features from a slide, first each patch is assigned to the closest prototype and then projected back to the patch coordinates in the original image to create a heatmap of prototype assignments. This creates a grey-level image where each pixel represents a single patch and the value is an integer indicating the prototype number the patch is most similar to, this grey-level image is identical to the prototype similarity maps seen in figures 6.2 and 6.3. From the grey-level image, the GLCM was calculated. This matrix is an $N$ by $N$ matrix where $N$ is the number of prototypes. Each cell in the matrix is the number of times a pixel with intensity value i is separated from another pixel with intensity value j at a particular distance k in a certain direction d. Here, the distance was set to 1 pixel and each of the 8 directions around the pixel were used. The matrix was then normalised to return the normalised co-occurrence matrix. From this matrix several statistics can be derived which provide information about the texture of the image. Here the contrast, homogeneity, angular second moment (ASM) and correlation were calculated. Welch's t-test was used to determine if there was any difference between these features depending on the slide label. The contrast feature measures the intensity contrast between a pixel and its neighbour over the whole image with a constant image having a contrast of 0. Homogeneity returns a value that measures the closeness of the distribution of elements in the GLCM to the diagonal of the GLCM. The ASM, also known as the energy, measures the sum of the squared elements in the GLCM with a value of 1 given for constant images. Correlation measures how correlated a pixel is to the neighbouring pixels over the whole image.

For Lung sub-type prediction all features except for homogeneity were significantly different between LUAD and LUSC cases ($P < 0.05$). LUAD cases had a higher mean contrast ($0.488 \pm 0.200$) compared to LUSC cases ($0.427 \pm 0.144$) and there was a significant difference (t(1051)=5.68, $p < 10^{-7}$). LUAD cases had a higher mean ASM ($0.280 \pm 0.097$) compared to LUSC cases ($0.262 \pm 0.074$) and there was a significant difference (t(1051)=3.45, $p < 10^{-3}$). LUAD cases also had a higher mean ASM ($0.280 \pm 0.097$) compared to LUSC cases ($0.262 \pm 0.074$) and there was a significant difference (t(1051)=3.45, $p < 10^{-3}$). LUAD cases had slightly lower mean correlation ($0.948 \pm 0.068$) compared to LUSC cases ($0.950 \pm 0.066$) and there was a significant difference (t(1051)=-5.26, $p < 10^{-6}$). However, there was not a significant difference between the homogeneity in either group (t(1051)=0.14, p=0.89). In addition, the percentage of counts of each prototype between LUAD and LUSC cases can be measured. LUSC cases had a higher average percentage of prototype 2 ($0.25 \pm 0.15$) compared with LUAD ($0.20 \pm 0.14$) and was significantly different (t(1051)=-4.48, $p < 10^{-5}$). LUAD cases had a slightly higher average percentage of prototype 1 ($0.15 \pm 0.14$) and prototype 8 ($0.11 \pm 0.11$) compared with LUSC cases, with both difference being significantly different, (t(1051)=4.63, $p < 10^{-5}$, t(1051)=4.60, $p < 10^{-5}$). The average count of all other prototypes was not found to be significantly different between LUAD and LUSC cases.

For breast cancer grading, a similar analysis was performed. None of the GLCM features were found to be statistically different between low/moderate and high grade cases using Welch's t-test (all p values $> 0.05$). This suggests that the textural appearance of prototype similarity maps are similar between low/moderate and high grade cases. As the lung dataset was much larger than the grading dataset it could be the case that more WSI of breast cancer would be needed to determine if there was a statistical difference. For counts of prototypes across a slide there was a significant difference is the percentage of prototypes 1, 2 and 5. These are for regions of adipose, dense tumour and stromal tissue. The most interesting factor here is that the average percentage of patches closest to prototype 2 in higher grade cases ($0.376 \pm 0.194$) is greater than in low/moderate grade cases ($0.260 \pm 0.175$) and was significantly different (t(161)=-3.65, $p < 10^{-3}$).

### 6.5.3 Ablation Study

To understand the importance of the different elements of Cross-Attention MIL an ablation study was performed. The first experiment investigates the contribution of the cluster loss and orthogonality loss on model performance and prototype discovery. The second experiment compares the cross-attention head against a prototype discovery step which discovers the prototypes through k-means clustering. Tables 6.4 and 6.5 show the results with and without the cluster loss and orthogonality loss on the test set for lung sub-typing and breast grading respectively. For consistency, the coefficients for both the cluster loss and orthogonality loss were set to a value of 1.0 and a total of 8 prototypes were used. From the results the best performance is achieved with the inclusion of both loss terms, especially for breast cancer grading where there is an 6.7% increase in accuracy and 5.2% increase in F1-score when using the additional loss terms. However, model performance is not vastly superior to the case where the losses are not included for the lung sub-typing dataset. Instead, the main benefit of the additional loss terms is in the model interpretability. Inclusion of the extra loss terms results in more diverse prototype discovery. To visualise this, the prototype similarity map, for a LUAD case, produced with and without the cluster loss and orthogonality loss can be seen in Fig 6.6. From the similarity maps it is clear that without the loss terms only two prototypes are found to be the closest to any of the patches within the image.

Table 6.6 shows the results using the k-means based prototype discovery step. In this step, prototype vectors are initially learned through k-means clustering instead of through model training. To learn the lung sub-typing prototype vectors, first 20 patches are randomly sampled from each slide in the training set and features are extracted at 10× magnification using a ResNet-50 pretrained on ImageNet. Then k-means is used to cluster the vectors using 8

Table 6.4: Ablation experiments for additional loss terms. Results are the average across 5 folds for LUAD vs LUSC classification using the test set.

| Method | AUC | Accuracy | F1-Score |
|---|---|---|---|
| Cross-Attention w/o losses | 0.932 ± 0.003 | 0.873 ± 0.011 | 0.873 ± 0.011 |
| Cross-Attention w/ Ortho | 0.931 ± 0.004 | 0.873 ± 0.014 | 0.873 ± 0.014 |
| Cross-Attention w/ Cls | **0.936** ± **0.009** | 0.868 ± 0.025 | 0.868 ± 0.025 |
| Cross-Attention w/ Cls & Ortho | 0.935 ± 0.012 | **0.876** ± **0.017** | **0.876** ± **0.016** |

Table 6.5: Ablation experiments for additional loss terms. Results are the average across 5 folds for low/moderate grade vs high breast cancer grade classification using the test set.

| Method | AUC | Accuracy | F1-Score |
|---|---|---|---|
| Cross-Attention w/o losses | 0.715 ± 0.022 | 0.685 ± 0.045 | 0.699 ± 0.037 |
| Cross-Attention w/ Ortho | 0.714 ± 0.018 | 0.6691 ± 0.035 | 0.704 ± 0.027 |
| Cross-Attention w/ Cls | 0.719 ± 0.018 | 0.739 ± 0.031 | 0.744 ± 0.029 |
| Cross-Attention w/ Cls & Ortho | **0.719** ± **0.014** | **0.752** ± **0.023** | **0.751** ± **0.024** |

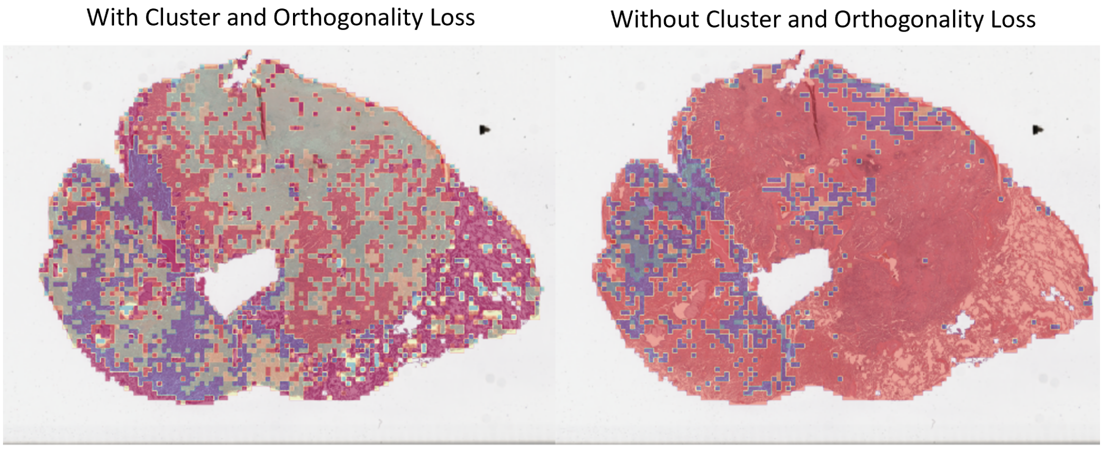

With Cluster and Orthogonality Loss
Without Cluster and Orthogonality Loss

Figure 6.6: Prototype similarity heatmaps with and without the cluster and orthogonality loss terms.

.

Table 6.6: Ablation experiments for k-means based prototype discovery step. Results are the average across 5 folds for Lung sub-typing and Breast Grading using the test set.

| Dataset | AUC | Accuracy | F1-Score |
|---|---|---|---|
| Lung Sub-type | 0.924 ± 0.013 | 0.875 ± 0.013 | 0.875 ± 0.013 |
| Breast Grading | 0.709 ± 0.024 | 0.691 ± 0.035 | 0.707 ± 0.032 |

centroids and these centroids are the new prototypes. To learn the grading prototype vectors, all patches are sampled from each slide in the training set which is possible here because the number of slides is much smaller. Then k-means is again used to cluster the vectors using 8 centroids which represent the prototypes. In both cases new prototypes are learned for each fold and the prototypes are then frozen during training so they do not update. An additional linear layer is also included with ReLU activation to transform the 1024-dimensional prototype vectors into 512-dimensions to match the dimensions of the key and value embeddings. The parameters of this linear layer are updated during training. Comparing the results in Table 6.6 to those from Tables 6.2 and 6.3, the use of the k-means based prototype discovery step results in a similar performance for lung cancer sub-typing. However, the end-to-end prototype learning seems to be beneficial over the k-means step when looking at breast cancer grading performance.

Overall, the addition of the cluster and orthogonality loss terms slightly increase performance but more importantly increase model interpretability and prototype diversity when learning prototypes end-to-end. The use of the k-means based prototype discovery step does produce similar performance for lung sub-typing but reduced performance for breast cancer grading. Furthermore, the prototype discovery step adds another stage to the approach whereas learning prototypes end-to-end incorporates it into the training process and can help reduce training times. The ablation study was performed using 10× magnification patches and results could be different when patches at different magnification are used.

## 6.6 Conclusion

In conclusion, a prototype based MIL approach has been proposed. This method learns a set of prototype vectors during the training process which each capture a different morphology with a WSI. A cross-attention head is also used to determine the similarity between each prototype and each patch within a slide which allows for a learnable similarity metric over something such as euclidean distance. The methodology was applied to both Lung cancer sub-typing and breast cancer grading. For the task of lung cancer sub-typing, performance was found to be comparable to other SOTA methods such as: ABMIL, CLAM and TransMIL. In the case of breast cancer grading, the proposed approach achieved superior performance compared to the other methods when using patches extracted at 10× magnification. An ablation study was also performed to determine the contribution of each element as well as a hyperparameter search to determine the optimal coefficients for each of the additional loss terms.

The proposed approach provides interpretability into the model decision by using an attention heatmap, similar to the other methods used for comparison. Additionally, the average global attention given to each prototype can be measured providing a global explanation instead of just a local one. Finally, the prototype similarity maps can be studied to understand textural features that are different between classes allowing for some understanding over the underlying biology without requiring detailed annotations.

A limitation of the current work is the lack of dependencies between patch embeddings. TransMIL utilises a transformer approach to update patch features based on the other patches within a slide. Incorporating a system like that into this approach could improve performance. Another limitation is that only one dataset was used for each task, and for the breast grading dataset specifically, the number of slides used was small. Future work would need to validate the approach on additional hold out test sets from other institutions to determine the robustness and generalisability of the proposed method. Although the method can use the prototype similarity maps to provide global explanations for class predictions in the form of textural and count based features, the differences are very subtle between classes. The features extracted from the GLCM were only found to be significantly different between LUAD and LUSC cases, for breast grading none of the features were significantly different. Contrast was the most different between LUAD and LUSC cases with an average difference of 0.061 however this was smaller than the standard deviation of contrast in both LUAD and LUSC cases. Therefore, although some insights can be gathered into the variation of tissue structures across a slide, they are often very subtle and would be hard to use for predicting for certain classes.

# Chapter 7

# Conclusion

## 7.1 Summary

This work ultimately aimed to explore interpretable AI methods for breast cancer WSI analysis which could provide insights into the decision making process and increase trust between the automated tool and trained pathologists. To summarise each of the chapters in this work:

Chapter 1 introduced the need for CAD tools to provide faster and more accurate cancer diagnostics, which has been made possible due to the transition towards digital pathology. This transition is discussed within this chapter along with the many benefits it has brought about. The challenges which need to be overcome to integrate successful CAD tools into the work environment are introduced with a focus on model explainability in order to overcome the "black-box" problem. The main contributions of this work were also highlighted.

In Chapter 2 the origins of cancer and the specific types of breast cancer were introduced. The methods for diagnosing and understanding the severity of a breast cancer case as well as the optimal treatment pathways were also discussed. Following this, automated cancer detection methods from the literature were presented to provide a background on the ways in which these methods have progressed over time. Hand-crafted features were first introduced including pixel, object and semantic level features. Following this, more data driven approaches which utilise DNN architectures were discussed. Supervised, weakly-supervised and unsupervised methodologies were introduced as well as their strength and limitations. Finally, the chapter introduced a taxonomy of explainability before discussing current SOTA methods for explainable AI and their applications in histopathology.

Chapter 3 describes the various datasets used throughout this work. Four datasets were used in total, of which three were open source. The fourth dataset was collected from the LTHT and provided both a patient's primary tumour slide and a lymph node slide. The methods for processing the datasets were explained. This involved a discussion of the rationale for including or removing patients from the dataset as well as the way in which slides were filtered to remove redundant information and effectively extract patches from each slide.

Chapter 4 investigated Prototypical Part Networks for interpretable diagnosis of both breast cancer grade and sub-type from ROI extracted from patients slides. The methodology takes a modern approach to case-based reasoning by providing examples from the dataset similar to a case in question. Overall, the prototype network performed similarly to the "black-box" counterparts while providing greater interpretability.

Chapter 5 looked at weakly-supervised methods for prediction of NPI group. ABMIL was used for both prediction of the individual parts that make up the index as well as for an end-to-end fusion approach which directly predicted the prognostic group. Prediction of the individual parts was found to be superior to the direct approach. The direct approach however has potential to be extended to tasks where only a single label is used for training (e.g grade or staging not available).

The final work in chapter 6 attempted to combine the prototypes with ABMIL to further increase the interpretability beyond heatmaps. Unlike similar methods, the method presented here learned prototypes during the training process which removed the need for a prototype discovery step. The learned set of prototypes provide a global explanation providing understanding of the important prototype per class. Additionally, the spatial arrangements of the prototype similarity maps can be evaluated, allowing for analysis of texture and prototype counts across a slide.

## 7.2 Contributions and Findings

The main contributions of the work were as follows:

1. Prototypical part networks were investigated for their effectiveness in analysis of ROIs from histopathology slides in chapter 4. Several modifications were made to the original implementation. These include the use of average pooling over max pooling, a cosine similarity metric and inclusion of an orthogonality loss term to increase prototype diversity. Using these modifications, the prototype part network was found to be superior to similar "black-box" models. The best performing Prototypical Part Network for breast cancer sub-typing used a ResNet-18 as the backbone, average pooling of the similarity scores and cosine distance to measure the similarity scores. This achieved a test AUC of $0.948 \pm 0.005$ and test ACC of $0.827 \pm 0.018$. Using the ResNet-18 backbone without the prototype layers achieved a test AUC of $0.963 \pm 0.019$ and test ACC of $0.775 \pm 0.071$. For breast cancer grading, the best Prototypical Part Network used a ResNet-18 backbone with average pooling of the similarity scores and cosine distance to measure the similarity scores. This configuration achieved a test AUC of $0.999 \pm 0.000$ and test ACC of $0.986 \pm 0.005$. A baseline ResNet-18 achieved a test AUC of $0.998 \pm 0.001$ and test ACC of $0.974 \pm 0.011$.

2. Weakly-supervised methods were used for prediction of NPI group in chapter 5. This is the first work to predict NPI group from WSI using automated methods. Two distinct approaches were investigated. The first approach aimed to predict the NPI group by first assessing the grade and lymph node stage from a patient's primary site and lymph node WSI. The grade and lymph node stage were combined with the pathologist reported invasive tumour size to predict the NPI group. The second approach aimed to predict the NPI group directly, without first determining grade, lymph node stage or invasive tumour size. This is a far more challenging problem but offers wider application for other tasks where multiple labels are not available or it would be time-consuming to produce them. Additionally, it can be used for patient outcome analysis where the relationships between bio-markers and outcomes is unknown. For example, for TNBC there is known relationship between low TILs-tumour and poor prognosis. However, in scenarios where a

relationship such as this is not known, the direct approach could be beneficial. Overall, it was found that the approach which predicted the individual index components was found to be the most accurate. For prediction into moderate-1, moderate-2 and poor/very poor groups, the best performing part prediction approach achieved a test class weighted F1-score of $0.822 \pm 0.033$ while the best performing direct prediction approach achieved a test F1-score of $0.611 \pm 0.078$. For prediction into a combined moderate-1 and moderate-2 class or poor/very poor class, the part prediction approach achieved a test f1-score of $0.789 \pm 0.069$ while the best performing direct prediction approach achieved a test f1-score of $0.760 \pm 0.028$.

3. A novel dataset which contains both a patient's primary site and lymph node image was used for prediction of NPI group. This is the first dataset used for model development which contains both types of slide images. The dataset was used to explore the potential of several fusion operators on top of an ABMIL backbone for WSI analysis. These fusion operators were fusion by vector concatenation, vector addition, average pooling, Hadamard-product and self-attention aggregation. The approaches used could be beneficial for additional tasks such as survival analysis where the inclusion of multiple patient slides could prove extremely beneficial.

4. A prototype based cross-attention MIL approach was introduced in chapter 6 which, unlike other similar methods, learned prototypes in an end-to-end manner. The approach provided local explanations through the use of attention heatmaps and global explanations by looking at the average attention given to each prototype. Additionally, this approach demonstrated that the spatial arrangements and counts of the prototype similarity maps could also be explored. Cross-attention MIL achieved an F1-score on the held out lung cancer sub-type prediction dataset of $0.875 \pm 0.019$ using patches extracted at 10× magnification. ABMIL achieved the second best F1-score of $0.873 \pm 0.020$ for the same task using patches at the same magnification. When looking at higher magnification of 20×, the best performing MIL method was ABMIL, achieving a F1-score of $0.869 \pm 0.012$. This was slightly higher than cross-attention MIL which achieved an F1-score of $0.857 \pm 0.007$. For the task of breast cancer grading, cross-attention MIL

achieved an F1-score of 0.761 ± 0.009 with ABMIL achieving the second higher score of 0.732 ± 0.047 when using 10× magnification patches. For patches at a magnification of 20×, CLAM achieved the highest F1-score of 0.686 ± 0.055 while cross-attention MIL got an F1-score of 0.673 ± 0.051. Overall, the prototype based cross-attention MIL method was found to perform similarly to other SOTA methods for lung cancer sub-typing but provided superior performance when grading breast cancer cases compared to current SOTA methods.

## 7.3 Limitations

There were several limitations present across the methods used throughout this work.

The first limitation is from the LTHT dataset used in both chapters 5 and 6. With only 163 patients post-processing, the dataset is relatively small which can limit model performance, as well as make it hard to draw conclusions about the effectiveness of the methods developed. The dataset also contains an unbalanced class distribution making it difficult to predict for some cases (Good prognosis, Very Poor prognosis and Grade 1). Finally, the dataset is only from one institution making it hard to understand how well the methods would generalise to other patient populations. On top of this, the datasets used in chapter 4 were also small with only 400 images in the BACH dataset and 300 in the breast grading dataset. Fortunately, because the images are not WSI they are easier to augment which helps to overcome this. To overcome this limitation, future work would need to compare the methods on larger datasets, preferably from separate institutions, however this is easier said than done as data can be time-consuming to collect.

Another limitation was the weakly supervised methods used in 5 and 6. ABMIL was used in both chapters and provides clear insights into the model decision making process by providing an attention heatmap explanation. However the method does not consider the relationship between patches within the slide. To overcome this limitation, more sophisticated methods such as graph neural networks or transformers could be used which might have produced better results. On the other hand, the dependencies between instances can make the model prediction harder to understand as the importance of a patch within a slide is dependent on the neighbouring patches as well. For example, transformers make use of self-attention layers which computes pairwise attention values between each patch in an image. To visualise Transformer models, these attentions can be considered as importance scores. This is usually done for a single attention layer. However, it is common for multiple layers to be stacked and simply averaging the attentions obtained for each patch, would lead to blurring of the signal [25] and would not consider the different roles of the layers. Additionally, graph neural networks update patch features through message passing between adjacent patches. Therefore, the importance of a patch is also based on the importance of neighbouring patches which adds another layer of complexity. Therefore, future work could build upon the findings here and look to incorporate patch dependencies and look at explainable AI methods for graphs [166] or transformers.

The interpretable methods explored and introduced in this work have not been tested in a clinical scenario. Therefore, it will be important for future research to determine the effectiveness of different XAI strategies to assist pathologists with a variety of tasks. Currently, the benefits of XAI methods have only been hypothesised and they should be effectively evaluated before being implements. Methods would need to be evaluated for their clarity, faith-fullness and detail. XAI methods should provide adequate information while being easily understood and accurately reflect the model decision process.

## 7.4 Future Work

In future work, the work presented in Chapter 5 could be extended to patient survival analysis. This would require a follow up on the same cohort of patients used in the initial study. This work would be beneficial for two reasons. The first is that it addresses a weakness with current survival analysis methods which only consider the primary tumour. The lymph nodes are extremely important for understanding breast cancer prognosis and it could be hypothesised that including them would improve model performance. Incorporating more lymph node slides per patient could also improve model performance by providing more data. Also, incorporating other lymph nodes other than the axillary lymph nodes, such as the internal mammary node would be interesting. However, this would come with it's own challenges, such are the best way to aggregate all this additional information so that key features are prioritised. The second benefit would be to assess the accuracy of the NPI by comparing the index against the patient outcomes. It would also be interesting to see if the direct prediction approach trained in 5 agrees with the original NPI group of the patient outcome data. Initial steps have been taken towards this by following up with the same cohort, however there are currently too few events for a study to be worthwhile at the present time.

Other future work could look to incorporate dependencies between patches. By doing this the features could be enriched by considering the surrounding slide morphologies. This could be done using graph neural networks or through self-attention, similar to transformers. More specifically, in 6 the final embeddings output from the cross-attention head could be passed through a self-attention head to consider the dependencies between the prototype features. This has the potential to improve model performance but would add in extra model parameters increasing computational and memory costs.

Although the methods presented here have made use of weakly-supervised methods in order to provide interpretable model predictions, the gold standard would be to provide model explanations in the form of human interpretable features. For this, tissue and cell types would need to be discovered across the slide. This would allow for identification of features such as cell densities, tissue ratios and cellular spatial arrangements to understand the biologies present. A model could then be fitted using these interpretable features allowing for knowledge distillation and greater human understanding. Some of the work in 6 began to look at this by considering the GLCM features of the prototype similarity maps and also looked at the counts of different prototypes. However, the prototypes discovered captured very broad tissue types such as tumour or adipose and did not directly predict for cells and their locations. Future work could therefore look at using annotated data to extract specific tissues and cells, leading to novel biomarkers which could be useful for various medical tasks.

# References

[1] Christian Abbet, Inti Zlobec, Behzad Bozorgtabar, and Jean Philippe Thiran. “Divide-and-Rule: Self-Supervised Learning for Survival Analysis in Colorectal Cancer”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 12265 LNCS. 2020, pp. 480–489.

[2] David Ahmedt-Aristizabal et al. “A survey on graph-based deep learning for computational histopathology”. In: *Computerized Medical Imaging and Graphics* 95 (2022), p. 102027.

[3] American Cancer Society. *Types of Breast Cancer — About Breast Cancer*. 2021.

[4] Mohamed Amgad et al. “NuCLS: A scalable crowdsourcing, deep learning approach and dataset for nucleus classification, localization and segmentation”. In: *GigaScience* 11 (Feb. 2021), giac037.

[5] Teresa Araujo et al. “Classification of breast cancer histology images using Convolutional Neural Networks”. In: *PLOS ONE* 12.6 (June 2017), e0177544.

[6] Guilherme Aresta et al. “BACH: Grand challenge on breast cancer histology images”. In: *Medical Image Analysis* 56 (Aug. 2019), pp. 122–139.

[7] Marc Aubreville et al. “Mitosis domain generalization in histopathology images — The MIDOG challenge”. In: *Medical Image Analysis* 84 (Feb. 2023), p. 102699.

[8] *Augmentor — Augmentor 0.2.9 documentation.*

[9] Sebastian Bach et al. “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation”. In: *PLoS ONE* 10.7 (July 2015), e0130140.

[10] Alina Jade Barnett et al. "IAIA-BL: A Case-based Interpretable Deep Learning Model for Classification of Mass Lesions in Digital Mammography". In: *Nature Machine Intelligence* 3.12 (Mar. 2021), pp. 1061–1070.

[11] Ajay Basavanhally et al. "Multi-field-of-view framework for distinguishing tumor grade in ER+ breast cancer from entire histopathology slides". In: *IEEE Transactions on Biomedical Engineering* 60.8 (2013), pp. 2089–2099.

[12] Babak Ehteshami Bejnordi et al. "Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images". In: *Journal of Medical Imaging* 4.4 (2017), pp. 044504–044504.

[13] Soufiane Belharbi et al. *Negative Evidence Matters in Interpretable Histology Image Classification.* Dec. 2022.

[14] Aïcha BenTaieb and Ghassan Hamarneh. "Predicting cancer with a recurrent visual attention model for histopathology images". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11.* Vol. 11071 LNCS. Springer Verlag, Sept. 2018, pp. 129–137.

[15] Kaustav Bera et al. "Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology". In: *Nature Reviews Clinical Oncology* 16.11 (Nov. 2019), pp. 703–715.

[16] Alexander Binder et al. "Morphological and molecular breast cancer profiling through explainable machine learning". In: *Nature Machine Intelligence* 3.4 (Apr. 2021), pp. 355–366.

[17] H. J. Bloom and W W Richardson. "Histological grading and prognosis in breast cancer a study of 1409 cases of which 359 have been followed for 15 years". In: *British Journal of Cancer* 11.3 (1957), pp. 359–377.

[18] Nadia Brancati, Giuseppe De Pietro, Daniel Riccio, and Maria Frucci. "Gigapixel Histopathological Image Analysis Using Attention-Based Neural Networks". In: *IEEE Access* 9 (2021), pp. 87552–87562.

[19] Freddie Bray et al. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: A Cancer Journal for Clinicians* 68.6 (2018), pp. 394–424.

[20] Gabriele Campanella et al. "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images". In: *Nature Medicine* 25.8 (Aug. 2019), pp. 1301–1309.

[21] Cancer Research UK. *Breast cancer statistics — Cancer Research UK.* 2014.

[22] Mathilde Caron et al. "Emerging Properties in Self-Supervised Vision Transformers". In: *Proceedings of the IEEE International Conference on Computer Vision.* Apr. 2021, pp. 9650–9660.

[23] Christine L Carter, Carol Allen, and Donald E Henson. "Relation of Tumor Size, Lymph Node Status, and Survival in 24,740 Breast Cancer Cases". In: *Cancer* 63.1 (1989), pp. 181–187.

[24] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks". In: *2018 IEEE winter conference on applications of computer vision (WACV).* IEEE, 2018, pp. 839–847.

[25] Hila Chefer, Shir Gur, and Lior Wolf. "Transformer Interpretability Beyond Attention Visualization". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* 2021, pp. 782–791.

[26] Chaofan Chen et al. "This looks like that: Deep learning for interpretable image recognition". In: *Advances in Neural Information Processing Systems.* Vol. 32. 2019.

[27] Hao Chen, Xi Wang, and Pheng Ann Heng. "Automated mitosis detection with deep regression networks". In: *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI).* Ed. by IEEE. Vol. 2016-June. IEEE Computer Society, June 2016, pp. 1204–1207.

[28] Richard J. Chen, Rahul G. Krishnan, Richard J. Chen, and Rahul G. Krishnan. "Self-Supervised Vision Transformers Learn Visual Concepts in Histopathology". In: *arXiv* (Mar. 2022), arXiv:2203.00585.

[29] Richard J. Chen et al. "Multimodal Co-Attention Transformer for Survival Prediction in Gigapixel Whole Slide Images". In: *Proceedings of the IEEE International Conference on Computer Vision.* 2021, pp. 4015–4025.

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

[31] Richard J. Chen et al. "Whole Slide Images are 2D Point Clouds: Context-Aware Survival Prediction Using Patch-Based Graph Convolutional Networks". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24* 12908 LNCS (2021), pp. 339–349.

[32] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. "A simple framework for contrastive learning of visual representations". In: *International conference on machine learning.* PMLR, 2020, pp. 1597–1607.

[33] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. "Improved Baselines with Momentum Contrastive Learning". In: *arXiv preprint arXiv:2003.04297* (2020).

[34] Hao Cheng et al. "Double Attention for Pathology Image Diagnosis Network with Visual Interpretability". In: *2020 International Joint Conference on Neural Networks (IJCNN).* IEEE, July 2020, pp. 1–8.

[35] Jianlin Cheng, Zheng Wang, and Gianluca Pollastri. "A neural network approach to ordinal regression". In: *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence).* IEEE, 2008, pp. 1279–1284.

[36] Ozan Ciga, Tony Xu, and Anne Louise Martel. "Self supervised contrastive learning for digital histopathology". In: *Machine Learning with Applications* 7 (Mar. 2022), p. 100198.

[37] Geoffrey M Cooper and Robert E. Hausman. "The Development and Causes of Cancer". In: *The Cell: A Molecular Approach* 2 (2000), pp. 719–728.

[38] Pierre Courtiol, Eric W Tramel, Marc Sanselme, and Gilles Wainrib. "Classification and Disease Localization in Histopathology Using Only Global Labels: A Weakly-Supervised Approach". In: *arXiv preprint arXiv:1802.02212* (2018).

[39] Pierre Courtiol et al. "Deep learning-based classification of mesothelioma improves prediction of patient outcome". In: *Nature Medicine* 25.10 (Oct. 2019), pp. 1519–1525.

[40] Heather D. Couture et al. "Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype". In: *npj Breast Cancer* 4.1 (Dec. 2018), p. 30.

[41] Angel Cruz-Roa et al. "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks". In: *Medical Imaging 2014: Digital Pathology.* Vol. 9041. 2014, p. 904103.

[42] Angel Cruz-Roa et al. "High-throughput adaptive sampling for whole-slide histopathology image analysis (HASHI) via convolutional neural networks: Application to invasive breast cancer detection". In: *PLoS ONE* 13.5 (May 2018), e0196828.

[43] Sarah Jane Dawson, Oscar M. Rueda, Samuel Aparicio, and Carlos Caldas. "A new genome-driven integrated classification of breast cancer and its implications". In: *EMBO Journal* 32.5 (Mar. 2013), pp. 617–628.

[44] Olivier Dehaene et al. "Self-Supervision Closes the Gap Between Weak and Strong Supervision in Histology". In: *arXiv preprint arXiv:2012.03583* (Dec. 2020).

[45] James A. Diao et al. "Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes". In: *Nature Communications* 12.1 (Dec. 2021), pp. 1–15.

[46] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. "Solving the multiple instance problem with axis-parallel rectangles". In: *Artificial Intelligence* 89.1-2 (Jan. 1997), pp. 31–71.

[47] Kosmas Dimitropoulos et al. "Grading of invasive breast carcinoma through Grassmannian VLAD encoding". In: *PLoS ONE* 12.9 (Sept. 2017), e0185110.

[48] Meidan Ding, Aiping Qu, Haiqin Zhong, and Hao Liang. "A Transformer-based Network for Pathology Image Classification". In: *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).* IEEE, Jan. 2021, pp. 2028–2034.

[49] Scott Doyle et al. "Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features". In: *2008 5th IEEE Inter-*

*national Symposium on Biomedical Imaging: From Nano to Macro, Proceedings, ISBI.* 2008, pp. 496–499.

[50] Chaitanya Dwivedi et al. "Multi stain graph fusion for multimodal integration in pathology". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* Vol. 2022-June. 2022, pp. 1835–1845.

[51] Babak Ehteshami Bejnordi et al. "Using deep convolutional neural networks to identify and classify tumor-associated stroma in diagnostic breast biopsies". In: *Modern Pathology* 31.10 (Oct. 2018), pp. 1502–1512.

[52] Ali Foroughi pour et al. "Deep learning features encode interpretable morphologies within histological images". In: *Scientific Reports* 12.1 (June 2022), p. 9428.

[53] Marcus H. Galea, Roger W. Blamey, Christopher E. Elston, and Ian O. Ellis. "The Nottingham prognostic index in primary breast cancer". In: *Breast Cancer Research and Treatment* 22 (1992), pp. 207–219.

[54] Paula S. Ginter et al. "Histologic Grading of Breast Carcinoma: A Multi-Institution Study of Interobserver Variation Using Virtual Microscopy". In: *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* 34.4 (Apr. 2021), p. 701.

[55] Ben Graham et al. "LeViT: a Vision Transformer in ConvNet's Clothing for Faster Inference". In: *Proceedings of the IEEE International Conference on Computer Vision* (2021), pp. 12259–12269.

[56] Simon Graham et al. "Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images". In: *Medical Image Analysis* 58 (Dec. 2019), p. 101563.

[57] M. Graziani, V. Andrearczyk, Marchand Maillet S., and H. Müller. "Concept attribution: Explaining CNN decisions to physicians". In: *Computers in Biology and Medicine* 123 (Aug. 2020), p. 103865.

[58] Corentin Gueréndel, Phil Arnold, and Ben Torben-Nielsen. "Creating small but meaningful representations of digital pathology images". In: *MICCAI 2021 Workshop Computational Pathology (COMPAY).* Vol. 156. 2021, pp. 206–215.

[59] Noriaki Hashimoto et al. "Multi-scale Domain-Adversarial Multiple-instance CNN for Cancer Subtype Classification with Unannotated Histopathological Images". In: *Pro-*

*ceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* 2020, pp. 3852–3861.

[60] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* Vol. 2016-Decem. 2016, pp. 770–778.

[61] Simin He et al. "Combining Deep Learning with Traditional Features for Classification and Segmentation of Pathological Images of Breast Cancer". In: *Proceedings - 2018 11th International Symposium on Computational Intelligence and Design, ISCID 2018.* Vol. 1. Institute of Electrical and Electronics Engineers Inc., July 2018, pp. 3–6.

[62] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (Nov. 1997), pp. 1735–1780.

[63] Andreas Holzinger et al. "Towards the Augmented Pathologist: Challenges of Explainable-AI in Digital Pathology". In: *arXiv preprint arXiv:1712.06657* (2017).

[64] Le Hou et al. "Patch-Based Convolutional Neural Network for Whole Slide Tissue Image Classification". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* Vol. 2016-Decem. 2016, pp. 2424–2433.

[65] Yongxiang Huang and Albert C.S. Chung. "Evidence localization for pathology images using weakly supervised learning". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).* Vol. 11764 LNCS. Springer, Oct. 2019, pp. 613–621.

[66] Ziwang Huang et al. "Integration of Patch Features Through Self-supervised Learning and Transformer for Survival Analysis on Whole Slide Images". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).* Vol. 12908 LNCS. Springer, Cham, Sept. 2021, pp. 561–570.

[67] Maximilian Ilse, Jakub M Tomczak, and Max Welling. "Attention-based deep multiple instance learning". In: *35th International Conference on Machine Learning, ICML 2018.* Vol. 5. 2018, pp. 3376–3391.

[68] Been Kim, Rajiv Khanna, and Oluwasanmi Koyejo. "Examples are not enough, learn to criticize! Criticism for interpretability". In: *Advances in Neural Information Processing Systems.* 2016, pp. 2288–2296.

[69] Been Kim et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)". In: *International conference on machine learning.* PMLR, 2018, pp. 2668–2677.

[70] Eunji Kim, Siwon Kim, Minji Seo, and Sungroh Yoon. "Xprotonet: Diagnosis in chest radiography with global and local explanations". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* 2021, pp. 15714–15723.

[71] Janet L Kolodner. "An Introduction to Case-Based Reasoning". In: *Artificial Intelligence Review* 6 (1992), pp. 3–34.

[72] Bin Kong et al. "Cancer metastasis detection via spatially structured deep network". In: *Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings 25.* Springer Verlag, 2017, pp. 236–248.

[73] Sonal Kothari, John H. Phan, Todd H. Stokes, and May D. Wang. "Pathology imaging informatics for quantitative analysis of whole-slide images". In: *Journal of the American Medical Informatics Association* 20.6 (Nov. 2013), pp. 1099–1108.

[74] M. Muthu Rama Krishnan et al. "Computer vision approach to morphometric feature analysis of basal cell nuclei for evaluating malignant potentiality of oral submucous fibrosis". In: *Journal of Medical Systems* 36.3 (2012), pp. 1745–1756.

[75] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet classification with deep convolutional neural networks". In: *Communications of the ACM* 60.6 (2017), pp. 84–90.

[76] Neeraj Kumar et al. "A Dataset and a Technique for Generalized Nuclear Segmentation for Computational Pathology". In: *IEEE transactions on medical imaging* 36.7 (2017), pp. 1550–1560.

[77] Narmin Ghaffari Laleh et al. "Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology". In: *Medical Image Analysis* (May 2022), p. 102474.

[78] Andrew H.S. Lee and Ian O. Ellis. “The Nottingham prognostic index for invasive carcinoma of the breast”. In: *Pathology and Oncology Research* 14 (June 2008), pp. 113–115.

[79] Bin Li, Yin Li, and Kevin 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.* 2021, pp. 14318–14328.

[80] Chunyuan Li, Xinliang Zhu, Jiawen Yao, and Junzhou Huang. “Hierarchical Transformer for Survival Prediction Using Multimodality Whole Slide Images and Genomics”. In: *Proceedings - International Conference on Pattern Recognition.* Vol. 2022-Augus. Institute of Electrical and Electronics Engineers Inc., 2022, pp. 4256–4262.

[81] Guangli Li et al. “Multi-view Attention-guided Multiple Instance Detection Network for Interpretable Breast Cancer Histopathological Image Diagnosis”. In: *IEEE Access* 9 (2021), pp. 79671–79684.

[82] Hang Li et al. “DT-MIL: Deformable Transformer for Multi-instance Learning on Histopathological Image”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24.* Vol. 12908 LNCS. Springer, 2021, pp. 206–216.

[83] Jiayun Li et al. “A multi-resolution model for histopathology image classification and localization with multiple instance learning”. In: *Computers in Biology and Medicine* 131 (2021), p. 104253.

[84] Jiayun Li et al. “An attention-based multi-resolution model for prostate whole slide image classification and localization”. In: *arXiv* (2019).

[85] Lian Tao Li, Guan Jiang, Qian Chen, and Jun Nian Zheng. “Predic Ki67 is a promising molecular target in the diagnosis of cancer (Review)”. In: *Molecular Medicine Reports* 11.3 (Mar. 2015), pp. 1566–1572.

[86] Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. “Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions”. In: *32nd AAAI Conference on Artificial Intelligence, AAAI 2018.* 2018, pp. 3530–3537.

[87] Ruoyu Li et al. “Graph CNN for survival analysis on whole slide pathological images”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 174–182.

[88] Yi Li and Wei Ping. “Cancer Metastasis Detection With Neural Conditional Random Field”. In: *arXiv* (2018).

[89] Yuqian Li, Junmin Wu, and Qisong Wu. “Classification of Breast Cancer Histology Images Using Multi-Size and Discriminative Patches Based on Deep Learning”. In: *IEEE Access* 7 (2019), pp. 21400–21408.

[90] Meiyan Liang et al. “Interpretable classification of pathology whole-slide images using attention based context-aware graph convolutional neural network”. In: *Computer Methods and Programs in Biomedicine* 229 (Feb. 2023), p. 107268.

[91] Geert Litjens et al. “Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis”. In: *Scientific Reports* 6.1 (May 2016), pp. 1–11.

[92] Huidong Liu and Tahsin Kurc. “Deep learning for survival analysis in breast cancer with whole slide image data”. In: *Bioinformatics* 38.14 (July 2022), pp. 3629–3637.

[93] Yun Liu et al. “Detecting Cancer Metastases on Gigapixel Pathology Images”. In: *arXiv preprint arXiv:1703.02442* (2017).

[94] Mengkang Lu et al. “SMILE : Sparse-Attention based Multiple Instance Contrastive Learning for Glioma Sub-Type Classification Using Pathological Images”. In: *MICCAI Computational Pathology (COMPAY) Workshop* 1 (2021), pp. 1–8.

[95] Ming Y Lu et al. *Data efficient and weakly supervised computational pathology on whole slide images.* 2020.

[96] Ming Y Lu et al. “Semi-Supervised Histology Classification using Deep Multiple Instance Learning and Contrastive Predictive Coding”. In: *arXiv preprint arXiv:1910.10825* (2019).

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

[98] Ming Y. Lu et al. “Data-efficient and weakly supervised computational pathology on whole-slide images”. In: *Nature Biomedical Engineering 2021 5:6* 5.6 (Mar. 2021), pp. 555–570.

[99] Scott M Lundberg and Su In Lee. “A unified approach to interpreting model predictions”. In: *Advances in Neural Information Processing Systems*. Vol. 2017-Decem. 2017, pp. 4766–4775.

[100] Weiming Mi et al. “Deep Learning-Based Multi-Class Classification of Breast Digital Pathology Images”. In: *Cancer Management and Research* Volume 13 (June 2021), pp. 4605–4617.

[101] Grégoire Montavon et al. “Explaining nonlinear classification decisions with deep Taylor decomposition”. In: *Pattern Recognition* 65 (May 2017), pp. 211–222.

[102] Ramin Nakhli et al. “Sparse Multi-Modal Graph Transformer with Shared-Context Processing for Representation Learning of Giga-pixel Images”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Mar. 2023, pp. 11547–11557.

[103] Loris Nannia, Stefano Ghidoni, and Sheryl Brahnam. “Ensemble of convolutional neural networks for bioimage classification”. In: *Applied Computing and Informatics* 17.1 (2018), pp. 19–35.

[104] Meike Nauta, Ron van Bree, and Christin Seifert. “Neural Prototype Trees for Interpretable Fine-grained Image Recognition”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2021, pp. 14928–14938.

[105] Peter Naylor, Marick Laé, Fabien Reyal, and Thomas Walter. “Segmentation of Nuclei in Histopathology Images by Deep Regression of the Distance Map”. In: *IEEE Transactions on Medical Imaging* 38.2 (Feb. 2019), pp. 448–459.

[106] Cam Nguyen, Zuhayr Asad, Ruining Deng, and Yuankai Huo. “Evaluating transformer-based semantic segmentation networks for pathological image segmentation”. In: *Medical Imaging 2022: Image Processing (Vol. 12032, pp. 942-947). SPIE.* 2022, p. 128.

[107] Erasmo Orrantia-Borunda et al. “Subtypes of Breast Cancer”. In: *Breast Cancer*. Exon Publications, Aug. 2022, pp. 31–42.

[108] Adam Paszke et al. *Automatic differentiation in pytorch*. 2017.

[109] Dino Pedreschi et al. “Meaningful Explanations of Black Box AI Decision Systems”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01 (July 2019), pp. 9780–9784.

[110] Vitali Petsiuk, Abir Das, and Kate Saenko. “RisE: Randomized input sampling for explanation of black-box models”. In: *arXiv preprint arXiv:1806.07421* (2018).

[111] Milda Pocevičiūtė, Gabriel Eilertsen, and Claes Lundström. “Survey of XAI in Digital Pathology”. In: *Artificial intelligence and machine learning for digital pathology: state-of-the-art and future challenges.* Vol. 12090 LNCS. Springer, 2020, pp. 56–88.

[112] Talha Qaiser and Nasir M. Rajpoot. “Learning Where to See: A Novel Attention Model for Automated Immunohistochemical Scoring”. In: *IEEE Transactions on Medical Imaging* 38.11 (Nov. 2019), pp. 2620–2631.

[113] Linhao Qu et al. “DGMIL: Distribution Guided Multiple Instance Learning for Whole Slide Image Classification”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part II (pp. 24-34).* June 2022, pp. 24–34.

[114] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “”Why should i trust you?” Explaining the predictions of any classifier”. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* Vol. 13-17-Augu. 2016, pp. 1135–1144.

[115] Mousumi Roy et al. “Convolutional autoencoder based model HistoCAE for segmentation of viable tumor regions in liver whole-slide images”. In: *Scientific Reports* 11.1 (Dec. 2021), p. 139.

[116] Dawid Rymarczyk et al. “Interpretable Image Classification with Differentiable Prototypes Assignment”. In: ().

[117] Dawid Rymarczyk et al. “ProtoMIL: Multiple Instance Learning with Prototypical Parts for Whole-Slide Image Classification”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases.* Springer Science and Business Media Deutschland GmbH, 2022, pp. 421–436.

[118] Ramprasaath R. Selvaraju et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization”. In: *International Journal of Computer Vision* 128.2 (Oct. 2018), pp. 336–359.

[119] Olcay Sertel et al. “Histopathological image analysis using model-based intermediate representations and color texture: Follicular lymphoma grading”. In: *Journal of Signal Processing Systems* 55.1-3 (2009), pp. 169–183.

[120] Muhammad Shaban et al. “A digital score of tumour-associated stroma infiltrating lymphocytes predicts survival in head and neck squamous cell carcinoma”. In: *The Journal of Pathology* 256.2 (Feb. 2022), pp. 174–185.

[121] Muhammad Shaban et al. “A Novel Digital Score for Abundance of Tumour Infiltrating Lymphocytes Predicts Disease Free Survival in Oral Squamous Cell Carcinoma”. In: *Scientific Reports* 9.1 (2019), p. 13341.

[122] Zhuchen Shao et al. “TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification”. In: *Advances in Neural Information Processing Systems* 34 (June 2021), pp. 2136–2147.

[123] Yash Sharma et al. “Cluster-to-Conquer: A Framework for End-to-End Multi-Instance Learning for Whole Slide Image Classification”. In: *Proceedings of Machine Learning Research*. PMLR, 2021, pp. 682–698.

[124] Yifan Shen et al. “Explainable Survival Analysis with Convolution-Involved Vision Transformer”. In: *Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022*. Vol. 36. 2. Association for the Advancement of Artificial Intelligence (AAAI), June 2022, pp. 2207–2215.

[125] Seo Jeong Shin et al. “Style transfer strategy for developing a generalizable deep learning application in digital pathology”. In: *Computer Methods and Programs in Biomedicine* 198 (Jan. 2021), p. 105815.

[126] Julio Silva-Rodriguez, Adrian Colomer, Jose Dolz, and Valery Naranjo. “Self-Learning for Weakly Supervised Gleason Grading of Local Patterns”. In: *IEEE Journal of Biomedical and Health Informatics* 25.8 (Aug. 2021), pp. 3094–3104.

[127] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. “Deep inside convolutional networks: Visualising image classification models and saliency maps”. In: *arXiv preprint arXiv:1312.6034* (2013).

[128] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *ICLR*. 2015.

[129] Jean F. Simpson et al. “Prognostic value of histologic grade and proliferative activity in axillary node-positive breast cancer: results from the Eastern Cooperative Oncology Group Companion Study, EST 4189”. In: *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 18.10 (2000), pp. 2059–2069.

[130] Gurmail Singh and Kin Choong Yow. “These do not look like Those: An interpretable deep learning model for image recognition”. In: *IEEE Access* 9 (2021), pp. 41482–41493.

[131] Mohamed Slaoui and Laurence Fiette. “Histopathology Procedures: From Tissue Sampling to Histopathological Evaluation”. In: *Drug Safety Evaluation: Methods and Protocols*. Vol. 691. Springer, 2011, pp. 69–82.

[132] Daniel Smilkov et al. “SmoothGrad: removing noise by adding noise”. In: *arXiv preprint arXiv:1706.03825* (2017).

[133] Leslie H Sobin, Mary K Gospodarowicz, and Christian Wittekind. *TNM classification of malignant tumours*. John Wiley & Sons, 2011.

[134] Iam Palatnik de Sousa, Marley Maria Bernardes Rebuzzi Vellasco, and Eduardo Costa da Silva. “Evolved Explainable Classifications for Lymph Node Metastases”. In: *Neural Networks* 148 (May 2020), pp. 1–12.

[135] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. “Striving for simplicity: The all convolutional net”. In: *ICLR*. 2015.

[136] Chetan L. Srinidhi, Ozan Ciga, and Anne L. Martel. “Deep neural network models for computational histopathology: A survey”. In: *Medical Image Analysis* 67 (Jan. 2021), p. 101813.

[137] Arunima Srivastava et al. “Imitating Pathologist Based Assessment With Interpretable and Context Based Neural Network Modeling of Histology Images”. In: *Biomedical Informatics Insights* 10 (Jan. 2018), p. 117822261880748.

[138] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. “Axiomatic attribution for deep networks”. In: *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.

[139] Christian Szegedy et al. “Going deeper with convolutions”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol. 07-12-June. IEEE Computer Society, Oct. 2015, pp. 1–9.

[140] Thomas E. Tavolara et al. “Automatic discovery of clinically interpretable imaging biomarkers for Mycobacterium tuberculosis supersusceptibility using deep learning”. In: *EBioMedicine* 62 (Dec. 2020), p. 103094.

[141] Naofumi Tomita et al. “Attention-Based Deep Neural Networks for Detection of Cancerous and Precancerous Esophagus Tissue on Histopathological Slides”. In: *JAMA network open* 2.11 (Nov. 2019), e1914645.

[142] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in Neural Information Processing Systems*. Vol. 2017-Decem. Neural information processing systems foundation, June 2017, pp. 5999–6009.

[143] Duc My Vo, Ngoc Quang Nguyen, and Sang Woong Lee. “Classification of breast cancer histology images using incremental boosting convolution networks”. In: *Information Sciences* 482 (May 2019), pp. 123–138.

[144] Wingates Voon et al. “Performance analysis of seven Convolutional Neural Networks (CNNs) with transfer learning for Invasive Ductal Carcinoma (IDC) grading in breast histopathological images”. In: *Scientific Reports 2022 12:1* 12.1 (Nov. 2022), pp. 1–19.

[145] Quoc Dang Vu, Kashif Rajpoot, Shan E.Ahmed Raza, and Nasir Rajpoot. “Handcrafted Histological Transformer (H2T): Unsupervised representation of whole slide images”. In: *Medical Image Analysis* 85 (Apr. 2023), p. 102743.

[146] Jiaqi Wang, Huafeng Liu, Xinyue Wang, and Liping Jing. “Interpretable Image Recognition by Constructing Transparent Embedding Space”. In: (2021).

[147] Shidan Wang et al. “Comprehensive analysis of lung cancer pathology images to discover tumor shape and boundary features that predict survival outcome”. In: *Scientific Reports* 8.1 (July 2018), pp. 1–9.

[148] Shujun Wang et al. “RMDL: Recalibrated multi-instance deep learning for whole slide gastric image classification”. In: *Medical Image Analysis* 58 (Dec. 2019), p. 101549.

[149] Xi Wang et al. “Weakly Supervised Deep Learning for Whole Slide Lung Cancer Image Analysis”. In: *IEEE Transactions on Cybernetics* 50.9 (2020), pp. 3950–3962.

[150] Xinggang Wang et al. “Revisiting multiple instance neural networks”. In: *Pattern Recognition* 74 (2018), pp. 15–24.

[151] Y. Wang et al. “Improved breast cancer histological grading using deep learning”. In: *Annals of Oncology* 33.1 (Jan. 2022), pp. 89–98.

[152] Yaqi Wang, Lingling Sun, Kaiqiang Ma, and Jiannan Fang. “Breast Cancer Microscope Image Classification Based on CNN with Image Deformation”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 10882 LNCS (2018), pp. 845–852.

[153] S. R. Wellings, H. M. Jensen, and R. G. Marcum. “An atlas of subgross pathology of the human breast with special reference to possible precancerous lesions”. In: *Journal of the National Cancer Institute* 55.2 (Aug. 1975), pp. 231–273.

[154] Suzanne C. Wetstein et al. “Deep learning-based breast cancer grading and survival analysis on whole-slide histopathology images”. In: *Scientific Reports* 12.1 (Sept. 2022), pp. 1–12.

[155] Ellery Wulczyn et al. “Deep learning-based survival prediction for multiple cancer types using histopathology images”. In: *PLoS ONE* 15.6 (June 2020), e0233678.

[156] Chensu Xie et al. “Beyond Classification: Whole Slide Tissue Histopathology Analysis By End-To-End Part Learning”. In: *Medical Imaging with Deep Learning*. PMLR, Sept. 2020, pp. 843–856.

[157] Yuanpu Xie et al. “Efficient and robust cell detection: A structured regression approach”. In: *Medical Image Analysis* 44 (Feb. 2018), pp. 245–254.

[158] Fuyong Xing et al. “Pixel-to-Pixel Learning with Weak Supervision for Single-Stage Nucleus Recognition in Ki67 Images”. In: *IEEE Transactions on Biomedical Engineering* 66.11 (Nov. 2019), pp. 3088–3097.

[159] Yunyang Xiong et al. “Nyströmformer: A Nyström-based Algorithm for Approximating Self-Attention”. In: *35th AAAI Conference on Artificial Intelligence, AAAI 2021*. Vol. 16. 16. Association for the Advancement of Artificial Intelligence, May 2021, pp. 14138–14148.

[160] Bolei Xu et al. “Look, investigate, and classify: A deep hybrid attention method for breast cancer classification”. In: *Proceedings - International Symposium on Biomedical Imaging*. Vol. 2019-April. IEEE Computer Society, Apr. 2019, pp. 914–918.

[161] Heechan Yang, Ji Ye Kim, Hyongsuk Kim, and Shyam P. Adhikari. "Guided Soft Attention Network for Classification of Breast Cancer Histopathology Images". In: *IEEE Transactions on Medical Imaging* 39.5 (May 2020), pp. 1306–1315.

[162] Pengshuai Yang et al. "A deep metric learning approach for histopathological image retrieval". In: *Methods* 179 (July 2020), pp. 14–25.

[163] Jiawen Yao et al. "Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks". In: *Medical Image Analysis* 65 (Oct. 2020), p. 101789.

[164] Feiyang Yu and Horace H.S. Ip. "Semantic content analysis and annotation of histological images". In: *Computers in Biology and Medicine* 38.6 (June 2008), pp. 635–649.

[165] Jin-Gang Yu et al. "Prototypical multiple instance learning for predicting lymph node metastasis of breast cancer from whole-slide pathological images". In: *Medical Image Analysis* 85 (Apr. 2023), p. 102748.

[166] Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. "Explainability in Graph Neural Networks: A Taxonomic Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.5 (2023), pp. 5782–5799.

[167] Cecilia Zappa and Shaker A. Mousa. "Non-small cell lung cancer: current treatment and future advances". In: *Translational Lung Cancer Research* 5.3 (June 2016), p. 288.

[168] Matthew D Zeiler and Rob Fergus. "Visualizing and understanding convolutional networks". In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13.* Vol. 8689 LNCS. PART 1. Springer, 2014, pp. 818–833.

[169] Hongrun Zhang et al. "DTFD-MIL: Double-Tier Feature Distillation Multiple Instance Learning for Histopathology Whole Slide Image Classification". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* Mar. 2022, pp. 18802–18812.

[170] Ji Zhang, Wynne Hsu Mong, and Li Lee. "Image Mining: Trends and Developments". In: *Journal of intelligent information systems* 19 (2002), pp. 7–23.

[171] Jingwei Zhang et al. "A joint spatial and magnification based attention framework for large scale histopathology classification". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2021, pp. 3776–3784.

[172] Michael Zhang, James Lucas, Jimmy Ba, and Geoffrey E. Hinton. "Lookahead Optimizer: k steps forward, 1 step back". In: *Advances in Neural Information Processing Systems* 32 (2019).

[173] Zizhao Zhang et al. "Pathologist-level interpretable whole-slide cancer diagnosis with deep learning". In: *Nature Machine Intelligence* 1.5 (May 2019), pp. 236–245.

[174] Yu Zhao et al. "Predicting Lymph Node Metastasis Using Histopathological Images Based on Multiple Instance Learning with Deep Graph Convolution". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* 2020, pp. 4836–4845.

[175] Yu Zhao et al. "SETMIL: Spatial Encoding Transformer-Based Multiple Instance Learning for Pathological Image Analysis". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, 2022, pp. 66–76.

[176] Bolei Zhou et al. "Learning Deep Features for Discriminative Localization". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* Vol. 2016-Decem. IEEE Computer Society, Dec. 2016, pp. 2921–2929.

[177] Bolei Zhou et al. "Object detectors emerge in deep scene CNNs". In: *ICLR.* 2015.

# Appendices

# Appendix A

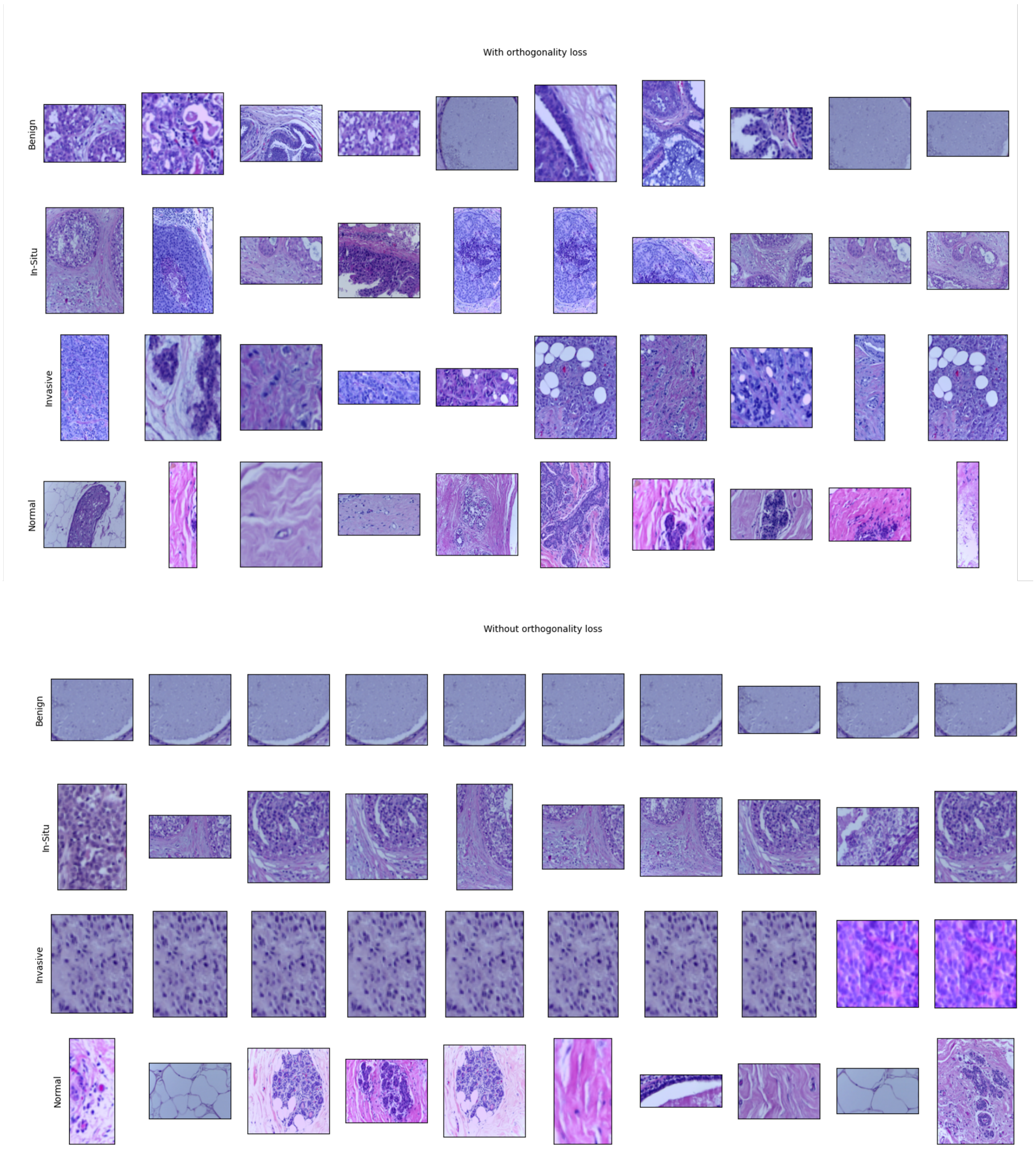


Figure A.1: Top and bottom show prototypes discovered when training with and without the orthogonality loss respectively. Each row shows the 10 prototypes discovered for each class on the BACH dataset.

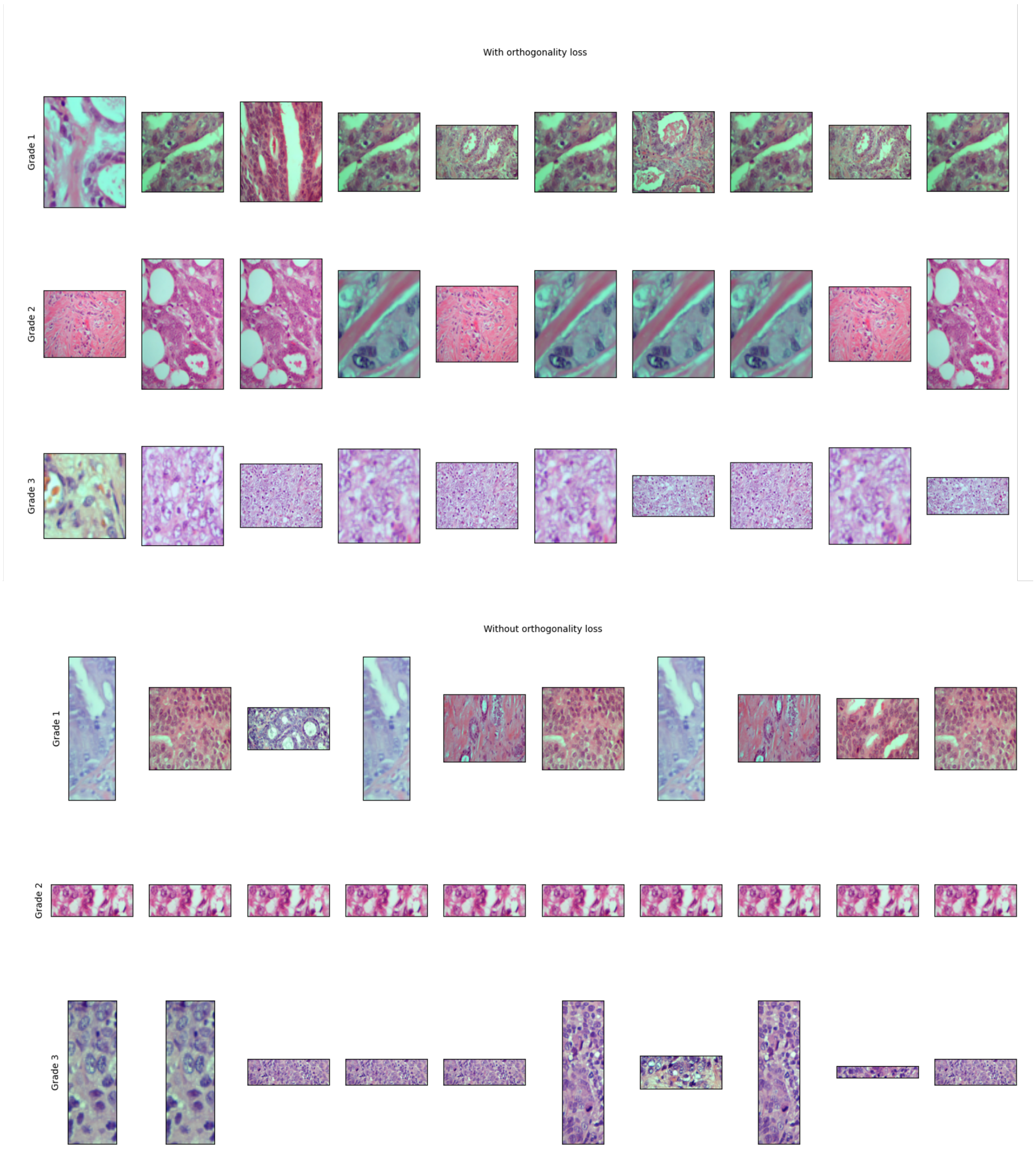


Figure A.2: Top and bottom show prototypes discovered when training with and without the orthogonality loss respectively. Each row shows the 10 prototypes discovered for each class on the Grading dataset.

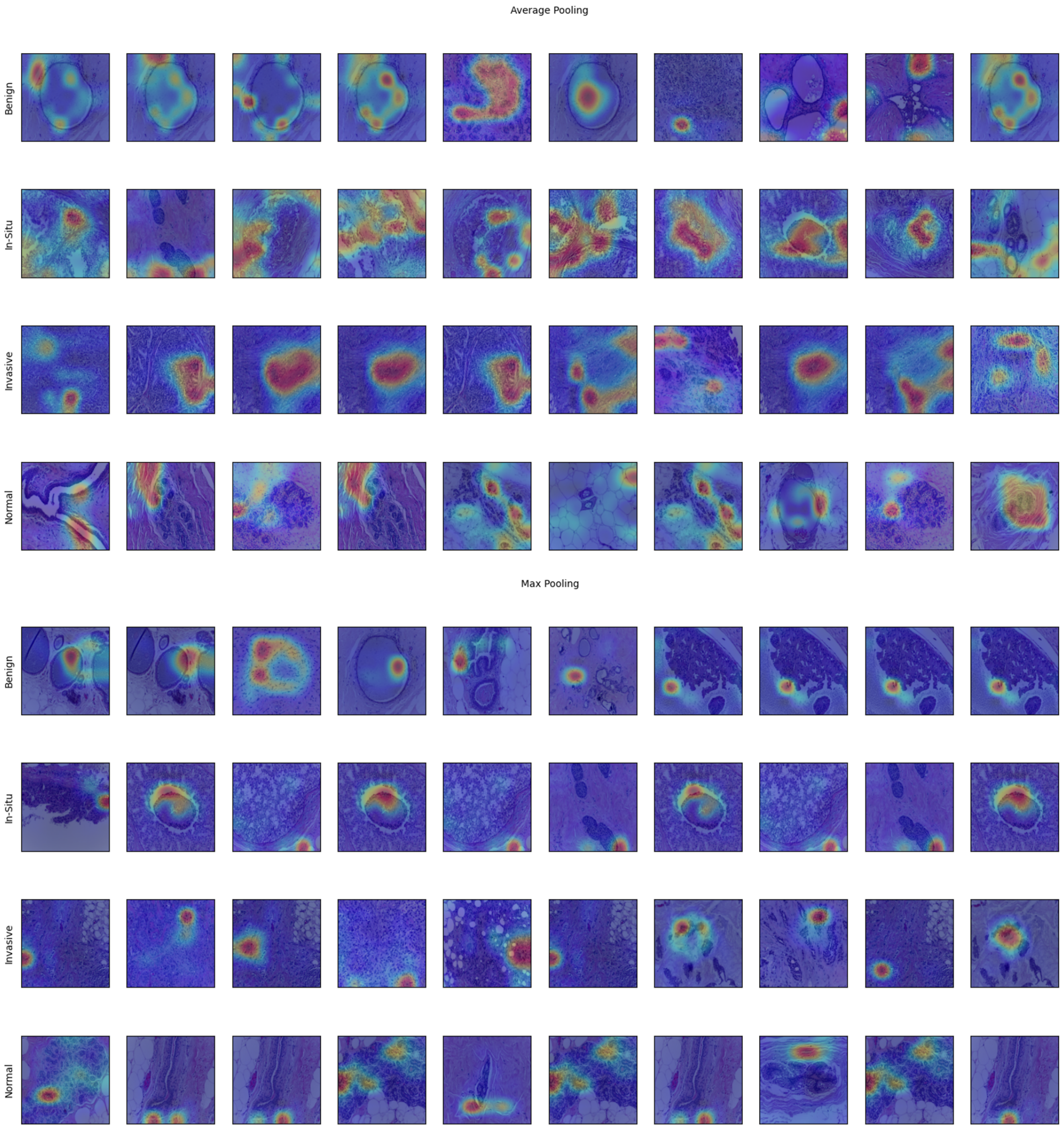


Figure A.3: Top and bottom show the prototype similarity maps when using average and max pooling respectively. Each row shows the 10 prototypes discovered for each class on the BACH dataset. Each image is the exact image used when visualising a specific prototype.

# Appendix B

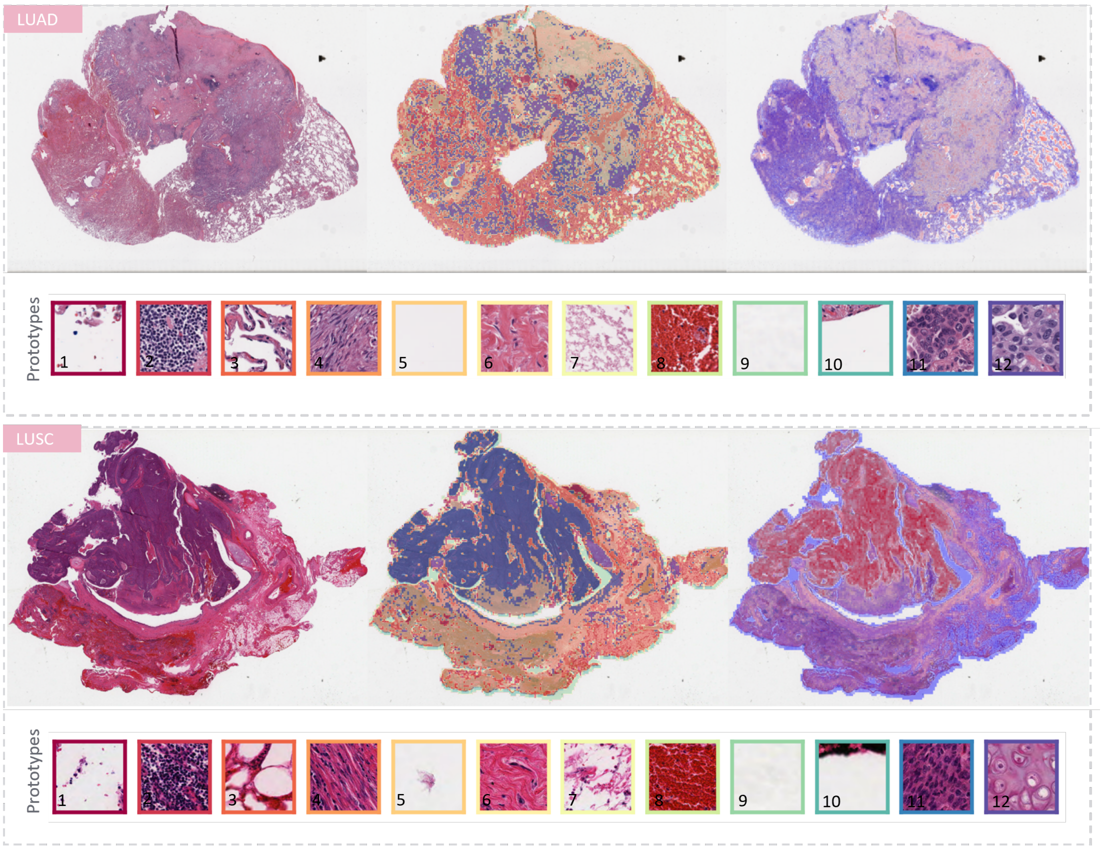


Figure B.1: Cross-Attention MIL attention maps for LUAD and LUSC cases using 20× patches. Top: (Left) Original WSI, (Middle) prototype similarity map, (Right) attention heatmap with higher attention in red and lower attention in blue. Bottom: this shows an example patch closest to each prototype.

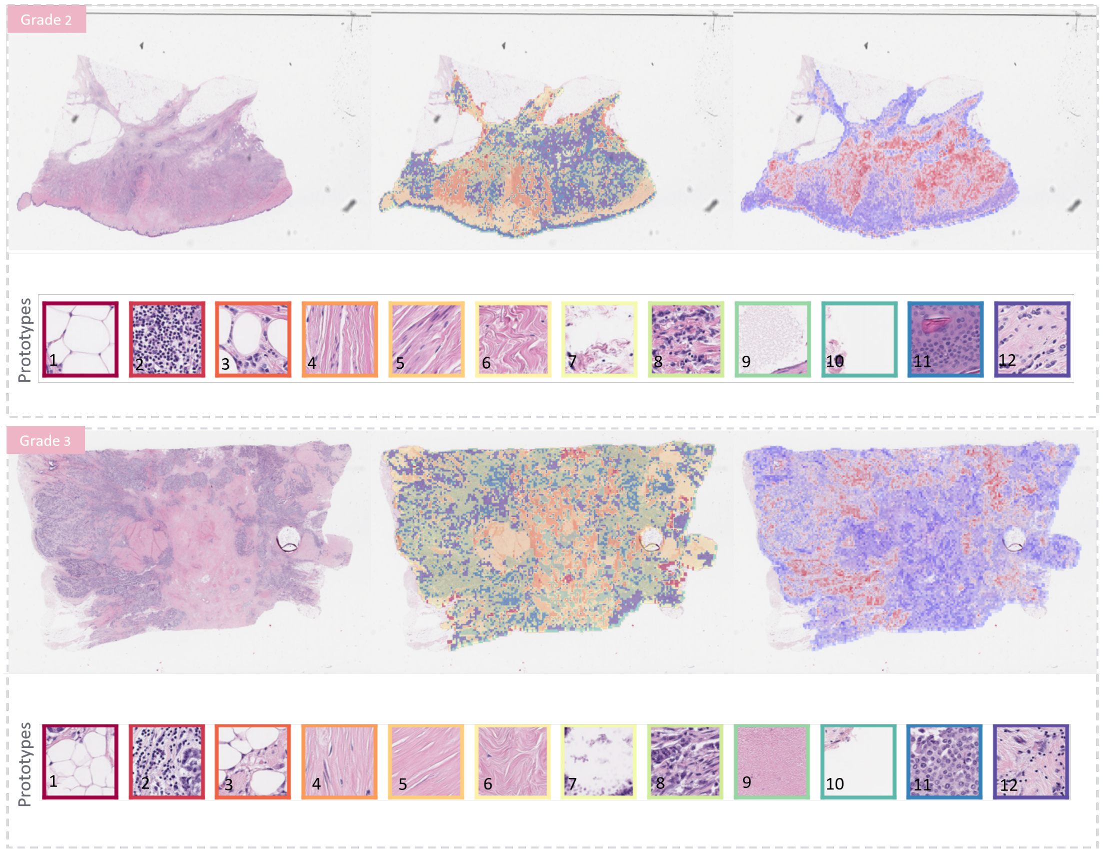


Figure B.2: Cross-Attention MIL attention maps for a grade 2 and grade 3 cases using 20× patches. Top: (Left) Original WSI, (Middle) prototype similarity map, (Right) attention heatmap with higher attention in red and lower attention in blue. Bottom: this shows an example patch closest to each prototype.