Evaluating Feature Extractors in Digital Pathology

Rithik Soni

May 2025

Abstract

Digital Pathology involves digitalising glass slides and analysing them using computational techniques, enabling scalable, reproducible, and high-throughput tissue analysis for both clinical diagnostics and research. However, whole slide images (WSIs) are extremely large—often exceeding one billion pixels leading to substantial high memory and computational demands. To address this, WSIs are typically divided into smaller non-overlapping segments called patches, which ease processing but introduce challenges such as loss of contextual information and reliance for robust, generalisable feature extractors. Recently, foundation models trained on large unlabelled datasets using self-supervised learning have shown potential in overcoming these limitations by learning transferable and annotation free representations of histological structures. Despite their promise, the performance of these models can vary considerably across datasets and tasks, including morphology classification, biomarker detection, and prognosis prediction.

In recent studies, author Wolfein in the papers [1] [2] benchmarked feature extractors to assess several key questions regarding stain normalisation and whether it remains a necessary preprocessing step, which feature extractors perform best for slide-level classification, and how magnification affects downstream performance. Wolflein concluded that stain normalisation was not necessary for weakly supervised tasks, significantly reducing computational memory usage. However, the study suggested that the most consequential choice for feature extraction is the specific extractor itself, having evaluated over 10,000 training runs across 14 models both with and without stain normalisation and examined the impact of image augmentations on feature extraction performance. This dissertation builds on this work by examining how image augmentations affect intermediate pathology tasks, particularly focusing on their impact on tumour segmentation.

1 Declaration

I hereby certify that this dissertation, which is approximately 9876 words in length, has been composed by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a degree. This project was conducted by me at the University of St Andrews from May 2025 to August 2025 towards fulfilment of the requirements of the University of St Andrews for the degree of MSc Artificial Intelligence under the supervision of Dr Ognjen Arandjelović . In submitting this project report to the University of St Andrews, I give permission for it to be made available for use in accordance with the regulations of the University Library. I

Contents

2 Plan

Date Range	Phase	Description
May 26 – June 6	Literature Review	Search through the literature and understand the processes and technology previously used to evaluate and assess feature extraction.
June 7 – June 15	Initial GPU Evaluation	Run the GPU against the ground truth, assess AUROC, F1 score, and accuracy scores.
June 16 – June 23	Vision-only Evaluation	Assess the vision-only feature extractors against the ground truth.
June 23 – June 27	Augmentation Evaluation	Evaluate vision-only feature extractors with augmentations (rotation, enlargement, stretching). Compare results with ground truth. Prepare findings for the interim demo.
June 28 – July 6	Fixes from Demo	Implement fixes and feedback from the interim demonstration.
July 7 – July 31	Vision and Vision-Language Evaluation	Evaluate vision-language and vision-only feature extractors against ground truth. Assess impact of augmentations and image compression.
August 1 – August 12	Final Write-up	Final report and buffer for corrections.

Feature extractors being assessed:

- CTransPath
- CONCH,
- CONCH1.5
- UNI,
- UNI2
- VIRCHOW2
- GIGAPATH
- H optimus 1
- Lunit-DINO
- ImageNet (ResNet)

Key Metrics being assessed:

- Area Under the Receiver Operating Characteristic curve (AUROC)

- Accuracy
- Intersection over Union
- F1 Score

Training models to be done on: TCGA-BRCA and TCGA-CRC The foundation models will be tested on:

- CPTAC-BRCA [3]
- CPTAC-CRC
- Camelyon17 [4]

3 Introduction

Precision medicine is an iterative approach to treatment that continuously refines patient stratification and therapeutic strategies based on real-time data [5]. A core enabler of this paradigm is digital pathology, which involves the digitisation of glass slides and their analysis using computational techniques. The term ”telepathology” introduced in 1986 by Ronald S. Weinstein is defined as ”a pathological diagnosis transmitted over a distance, together with specific digital images of micro- and macroscopic preparations, clinical data, and information on cases sent to a pathologist via data link.” [6].This enables scalable, reproducible, and high-throughput tissue analysis, thereby enhancing diagnostic accuracy and supporting data-driven clinical decision-making [7].

A central element of digital pathology is the analysis of whole slide images (WSIs), which are extremely large-often exceeding one billion pixels- and present significant computational and memory challenges. To address these, WSIs are typically divided into smaller non-overlapping segments known as patches. While patch-based processing facilitates deep learning applications, it also introduces problems such as the loss of contextual information and dependence on robust, generalisable feature extractors.

Recent advances in foundation models trained with self-supervised learning on large unlabelled datasets have shown promise in addressing these issues. These models can learn transferable, annotation-free representations of histological structures. However, their performance varies widely across tasks like morphology classification,biomarker detection, and prognosis prediction. A recent study by Wolfein et al. [1] highlighted key questions about stain normalisation, the optimal choice of feature extractor, and the influence of magnification on downstream performance. The findings suggest that stain normalisation may not be necessary for weakly supervised learning tasks- significantly reducing computational overhead- but emphasis that the feature extractor itself has the most substantial impact on performance.

This dissertation builds on these insights by evaluating how image augmentations affect intermediate pathology tasks, particularly focusing on tumour segmentation. It also reviews the development digital pathology, its technology and the current practices used.

4 Literature Review

This literature review critically examines the evolution of digital pathology and the multifaceted challenges associated with Whole Slide Image (WSI) analysis, in terms of both computational and infrastructure demands. It focuses on the major barriers to effective implementation: the limited availability of diverse annotated datasets, the immense size and resolution of WSIs, and the complexity of extracting meaningful features from patch-based representations. The review further explores how recent advances in artificial intelligence—particularly the emergence of self-supervised and foundation models—are beginning to address these issues. By evaluating the role of feature extractors in enabling scalable, generalisable, and context-aware analysis, this review highlights the rapidly evolving landscape of digital pathology and its ongoing transformation.

4.1 The History and Growth of Digital Pathology

The term telepathology was first introduced by R.S Weinstein in his paper "Prospects for Telepathology," where he envisioned the remote transmission of pathology images to enable diagnosis at a distance. This early vision laid the foundation for what would later become a transformative field in modern diagnostics.

The 1990s saw significant progress, with the development of the first virtual microscope. Originally designed for Earth Science research, data management software was developed to analyse, visualise, and query satellite sensor data and assign it to geospatial regions. During a National Science Foundation Challenge, the Saltz group recognised that this software could be repurposed, and adapted for pathology. They repurposed this software to handle high- resolution, data intensive whole slide images (WSIs) produced through digitisation. Joel Saltz, et al. proposed the first virtual microscope [8] to retrieve, store and process digitised slides and "emulate the usual behaviour of a physical microscope". The virtual microscope implemented four basic functionality features to emulate a light microscope:

1. Local browsing to observe the region around the current view,
2. Fast browsing to locate an area of interest
3. Adjusting the magnification
4. Changing the focal plane.

As Liron Pantanowitz noted in their paper recapping the last 20 years of pathology: "The realization that this nascent technology,being developed to process satellite data, could be applied to WSIs was the nidus that triggered the development of software necessary to support the earliest Virtual Microscope that was developed from 1996 to 1998." [9] In parallel, automated commercial innovation was underway to accompany the static image development. In 1994, James Bascus of Bascus Labratories Inc. developed the first commercial slide scanner, known as BLISS (Bascus Labratories Inc. Slide Scanner).Costing $300,000 dollars to make. The BLISS was able to take a scan of a slide in 24 hours. This advancement marked the beginning of virtual micropsy, enabling high resolution slide digitisation for clinical and research use.

The early 2000s continued these development trends with the automated cellular imaging system

(ACIS). The ACIS system was designed to provide quantitative analysis, detect rare events and count objects. Most modern WSI systems contain two components a scanner that handles the image acquisition and a work station that handles the processing of the image.

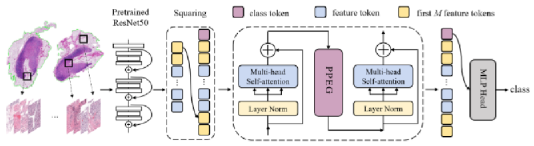


Figure 1: Key Components of a whole slide imaging system
[9]

Until 2009, there was no formal regulations governing the manufacturing of whole slide scanners and related devices. To address this gap, a United States Food and Drug Administration (FDA) advisory panel proposed that whole slide imaging systems used for primary clinical diagnosis in human tissue be classified as Class III medical devices. This classification is typically reserved for devices with high risks associated with them, introduced stringent regulatory standards and required manufacturers to demonstrate a high level of technical precision,validation, and safety before approval. The aim was to ensure a consistent quality and reliability in devices intended for critical diagnostic use. [10]

By the late 2000s through the early 2020s, technological advancements significantly bridge the gap that existed since the inception of telepathology. Current digital pathology methods have improved to the point where this is little perceptible difference between physical and digital pathology samples , Peter Caie stated:

> *"Humans are as adept at reporting from digital pathology samples as they are from glass mounted-ones."* [11]

Together, these developments catalysed the emergence of digital pathology, setting the stage for its integration into clinical workflows and the eventual convergence with deep learning methods.

The growth on digital pathology has had a positive impact on health care, the digitisation of pathology slides has eliminated the high costs associated with storing glass slides, and allows for remote pathological analysis leading to faster diagnoses times without a significant drop of in accuracy of the pathologist's results. While manual WSI image analysis is time consuming, the advances in both machine learning and deep learning algorithms the time can be reduced significantly. However, these methods are not flawless and new issues are introduced. [12]

4.2 Challenges in Whole Slide Image Analysis

One of the limitations in current WSI-based deep learning models is the lack of diverse annotated datasets. Many models have previously been trained on the CAMELYON16 dataset which consists of 400 WSIs focused on lymph node metastasis detection. While models perform well on this dataset, they often fail to generalise to real-world clinical settings where variations in staining

protocols, scanner types, and tissue morphology are more pronounced. Stain variation, used by To overcome this limitation, Campanella et al. [13] introduced a larger and more diverse dataset comprising over 12,000 prostate needle biopsy slides. Their study demonstrated that with increased dataset diversity and size, convolutional neural networks could achieve significantly higher performance. Prostate data was chosen for its medical relevance and computational difficulty, highlighting the importance of task-specific dataset design in digital pathology.

The lack of available, diverse data has lead to a generalisation of trained models. While there has been a trend emerging with more data publication in recent years, most of the data is still trained on the CAMELYON16 dataset.

Another critical challenge stems from the sheer size of WSIs. With dimensions reaching 100,000 by 100,000 pixels and file sizes exceeding 10 GB, WSIs place substantial demands on memory and processing power. This has traditionally constrained the adoption of deep learning methods despite their advantages. To mitigate this, two common approaches have emerged:

- **Downsampling**: Reducing image resolution to a manageable scale, typically between 10,000 and 20,000 pixels.
- **Patch-based Processing**: Dividing WSIs into smaller, non-overlapping patches that can be analysed. This method preserves local detail and enables parallel processing but introduces new challenges in aggregating local predictions to form a holistic slide-level interpretation.

Patch-based analysis involves dividing WSIs into smaller subsections of the large slide and assigning slide-level labels to each patch. While this approach is effective in condensing large gigapixel images, it introduces several limitations. A key concern is label inaccuracy. This is when a slide containing cancerous tissue is labelled positive, even though many of the patches contain normal tissue. This weak supervision can lead to instances when the model is trained and patches without cancerous tissue may be learned as positive meaning the result is a false positive. Another drawback is the loss of spatial context. By, analysing patches independently, the models are not learning of the information regarding the broader tissue architecture and spatial relationship between regions. The resulting spatial disconnect hampers tasks that rely on structural continuity. The main causes of spatial loss include:

- **Lack of global context:** No representation of the slide-wide tissue layout. This means large scale tissue patterns and arrangements such as tumour boundaries are not represented.
- **Discarding Overlaps** or **Positional metadata**: While patches are sampled from specific regions on the slide, the positional metadata is typically discarded during training. As a result of this, the model cannot infer spatial continuity or the potion relative to other patches.
- **Edge Effects:** Patches located at the edge of a slide may only capture tissue content leading the incomplete or missing features.
- **Isolated Feature Interpretation:** Each patch is processed in isolation without modelling spatial or contextual relationships with surrounding patches. This is an issue when nearby patches share tissue structure or pathological features which are relevant for accurate interpretation of the tissue.

4.3 Addressing the Problems

4.3.1 Downsampling

One of the proposed solutions to the large image sizes of WSIs is downsampling. Compression methods such as: Joint Photographic Experts Group (JPEG), or JPEG 2000 and Lempel-Ziv-Welch (LZW) have been found to reduce the file size by a factor of seven or more [14]. This reduction in file size allows the slides to be more easily stored and analysed. However, a key concern with these compression methods - particularly lossy compression is the irreversible loss of image information. Since data is discarded to achieve smaller file sizes, some fine details cannot be recovered. The study by [15] has shown that while lossy compression preserves morphologic features such as shape, structure, and spatial relationships relatively well, it significantly affects densitometric assessments. These assessments rely on the subtle intensity and colour variations, become less reliable due to compression artifacts. An alternative method to reduce image size is to discard blank regions of the slide. A blank region of a slide is an area of the slide that contains no tissue or diagnostically relevant content .This strategy is effective in reducing scanning time and computation requirements. Another factor to consider is the resolution and magnification of the WSI. Studies have shown that the relationship between image scale and field of view is intrinsic: as the field of view increases, the effective resolution is inherently limited by the display medium. This means that the final details are not visible regardless of the source image resolution. Conversely, at higher image resolutions, only a smaller region can physically be displayed on the screen at once due to spatial constraints. As a result, the amount of visible tissue and the level of detail are fundamentally linked where increasing one necessarily reduces the other.

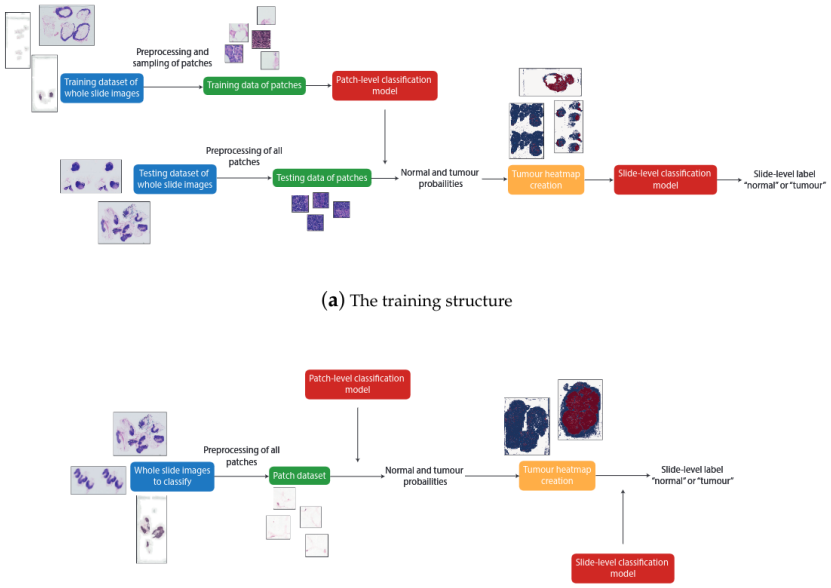
4.3.2 Patch-Based WSI analysis

As discussed earlier, an alternative approach to WSI analysis is patch based analysis, which involves dividing WSIs into smaller, fixed size patches that serve as input samples for training neural networks. This method is often the preferred method over downsampling, as it retains the high resolution details necessary for tissue analysis. Patch-based workflows consist of two main stages:

- **Feature Extraction:** Obtaining feature vectors for each patch using a pre-trained or fine tuned model.
- **Feature Aggregation:** Combining these patch-level features to produce a slide-level prediction,often through pooling or attention based mechanisms.

Both of these steps are parametrised using a neural network. [1]. A critical component of patch-based analysis is preprocessing, which prepares the data for consistent and efficient learning. Preprocessing usually involves the following four key steps [12].:

- **Tissue Segmentation**: which detects an blurry or unwanted areas. The areas in question are usually large regions that are irrelevant to the analysis of the tissue. This step usually saves computation power.
- **Colour Normalisation:** This alters the distribution of colour values in an image to standardise the range of colour in WSIs to ensure that only the relevant colour range appears in the analysis. This is an essential step as it reduces the stain variation in the images which can



(**a**) The training structure

Figure 2: Typical workflow of patch based classification [12]

cause bias. (More methods to combat stain normalisation are discussed and elaborated on in later sections).

- **Patch Extraction:** This involves taking square patches, typically $256 * 256$ pixels in size from the WSI for patch-based analysis.
- **Data Augmentation:** is the transformation of training data to new training data. This also prevents overfitting in the training data.

This patch-based approach has become a foundational workflow in computer pathology due to its scalability, compatibility with deep learning models, and ability to preserve high-resolution detail for downstream tasks.

The figure below illustrates the typical WSI workflow:

4.4 Current Deep Learning Methods for Patch-Based Analysis

The advancement of machine learning methods has been critical to the advancement of automated systems. Machine learning approaches can be split into multiple categories: feature engineering, weakly supervised methods, strongly supervised methods and unsupervised learning. Typically, deep learning methods rely on data that has been labelled with the target prediction, and systems learn to extract the relevant features from larger training data sets. Deep learning methods rely on data directly labelled with the target prediction. There are three typical ways in which deep learning methods are used for WSIs: for localisation such as tumour segmentation and detection and for segmentation such as tumour grading. [16]

4.4.1 Strongly Supervised Learning

A major limiting factor in image analysis for WSIs is the lack of training data. WSIs can be labelled at multiple layers including: pixel-level, patch-level-slide-level, lesion-level, and patient level. The most ideal scenario is to use a slide containing patch-level annotations to compare those to the experts. However, labelling patches is often time consuming and inefficient. This limitation hinders the use of strongly supervised learning methods as the This can be a time consuming and very expensive.

4.4.2 Weakly Supervised Learning

Weakly supervised learning methods have been adapted to address the problem of a lack of context in image classification by automatically extracting the refined valuable information from coarse labelled patches [17] [18].

4.4.3 Multiple Instance Learning

Multiple Instance Learning (MIL) is a form of supervised learning in which labels are assigned to sets of instances or "bags", rather than individual instances themselves. In practice,this entails assigning a single slide-level label to a collection of image patches extracted from a WSI. The goal is to train models that can infer bag-level labels under weak supervision, eliminating the need for expensive,fine-grained annotations. [19]

Traditional MIL approaches assume independence among patches and disregard spatial arrangement, a significant drawback for pathology images where spatial context is diagnostic. To address this, several studies have proposed integrating attention mechanisms or spatial context into MIL frameworks. These enhancements allow models to focus on the most informative regions within a slide,improving both predictive accuracy and interpretability in tasks like tumour classification. Initial MIL algorithms were primarily designed for binary classification problems, with the assumption that a bag is positive if at least one instance is positive. This assumption, often implemented with via max pooling, is overly rigid and unsuitable for multi-class classification. Consequently, more flexible aggregation strategies such as attention-based pooling have been proposed to more accurately reflect instance-level contributions.

Binary classification can defined as seen below:

$$Y_i = \begin{cases} 0, & \text{iff } \sum y_{i,j} = 0 \\ 1, & \text{otherwise} \end{cases} \qquad y_{i,j} \in \{0, 1\},\ j = 1 \dots n \tag{1}$$

$$\hat{Y_i} = S(\mathbf{X}_i) \tag{2}$$

This formulation enforces a strong assumption where the presence of a single positive instance defines the bag label.

Carbonneau et al. conducted a comprehensive survey [20] of MIL challenges and emphasised that limited understanding the problem characteristics has constrained the development and evaluation of MIL algorithms. They highlighted that the choice of dataset and problem formulation can

significantly impact experimental results, particularly when synthetic data fails to generalised to reflect real-world complexity. Their work categorises MIL problem characteristics into four key regions:

- **Prediction Level** – whether labels are predicted at the bag level or instance level.
- **Bag Composition** – how instances within a bag relate to the bag label.
- **Label Ambiguity** – the uncertainty in mapping instance labels to bag labels.
- **Data Distribution** – the underlying distribution of instances across and within bags.

Algorithms often excel in one type of prediction tasks (bag or instance level) but may struggle with the other, emphasising the need for approaches tailored to specific MIL scenarios.

4.4.4 Attention Based MIL

Maximilian Illse et al. [21] proposed a method, combining MIL with attention based mechanisms. This approach aimed to enhance the flexibility of MIL while preserving its interpretability. This method proposed the use of trainable attention mechanisms along with a Bernoulli distribution to model each label assignment for each patch and trained the model using the log likelihood. Another key factor in this mode is the replacement of standard permutation-invariant pooling operations such as mean or max pooling with a trainable weight. The weights are generated by a two-layer neural network, which implements the attention mechanism. This attention weight plays a role in identifying the most informative patches within each bag, thereby improving both performance and interpretability in MIL-based tasks. Illse et. al [21] defined the attention based mechanisms as the following:

$$z = \sum_{k=1}^{K} a_k h_k \tag{3}$$

where the attention weights are given by:

$$a_k = \frac{\exp\left(w^\top \tanh(V h_k)\right)}{\sum_{j=1}^{K} \exp\left(w^\top \tanh(V h_j)\right)} \tag{4}$$

The difference between the attention mechanism and previous MIL mechanisms is that with attention all instances are sequentially dependent in contrast whereas in this method assumes at all methods are independent.

4.4.5 Dual-stream Multiple Instance Learning

Li et al. [22] prosed a method called Dual stream multiple instance learning (DSMIL) to address the challenge of WSI classification using only slide-level labels, without the need for localised annotations. DSMIL is designed to perform both WSI-level classification and tumour localisation in a weakly supervised manner.

The proposed method combines the strength of two MIL paradigms: instance-based and embedding-based by utilising a dual-stream architecture. In traditional MIL:

- **Instance-based MIL**: A classifier predicts instance-level scores, and a pooling function (max-pooling) aggregates these scores to produce a bag-level prediction.
- **Embedding-based MIL**: A feature extractor maps each instance to an embedding, which is then aggregated to form a bag embedding. A classifier then predicts the bag label from this embedding.

Each approach has its strengths: instance-based methods are better at identifying key instances like tumour regions, while embedding-based methods often achieve higher accuracy but they are less interpretable. DSMIL jointly learns both streams to benefit the strengths of each.

The dual-stream consists of:

1. **Instance-level stream**: Applies an instance classifier with max-pooling to detect the most relevant patches (key instances) contributing to the bag label.
2. **Embedding-level stream**: Aggregates instance embeddings to form a holistic bag representation for classification
3. **Joint scoring**: Final predictions are computed by combining the instance-level and embedding-level scores.

Additionally, DSMIL employs a pyramidal multi-scale attention strategy to capture both local and global context within WSIs. It also incorporates self-supervised contrastive learning for each feature extraction which improves representation quality and reduces memory consumption during training. By combining both granular (instance-level) and holistic (embedding-level) perspectives, DSMIL offers a powerful framework for weakly supervised learning in digital pathology.

4.4.6 Transformer based Correlated MIL

Previous methods of MIL are based on the assumption that all instances in a bag are independent and identically distributed. However this assumption is not always valid. In many cases, pathologists often consider the contextual information around a single area and the correlation information between different areas when making a diagnostic decision. So it is important that this context is considered. With this in mind, authors Zhuchen Shao et al. proposed a Transformer based Correlated MIL method for WSI classification. [23]. This method used a transformer which had been used in several vision tasks such as end to end object detection [24] and as encoders for medical image segmentation [25](case studies) due to its ability to describe the correlation between different segments or tokens and model long distance information. However, traditional transformer sequences are limited by their computational complexity and can only compute shorter sequences, therefore not suitable for large size images such as WSIs. This paper seeks to leverage the transformers and MIL and thus this paper proposes TransMIL.

The authors proposed a token pyramid transformer module with two transformer layers and a positional encoding layer, where transformer layers are designed for aggregating morphological information and pyramid position encoding generator (PPEG) which is designed for encoding spatial information.

To address long instances problems in WSIs, the self-attention mechanism is optimised using the

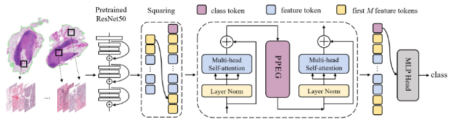


Figure 3: TransMil Architecture [23]

Nystrom method to approximate self-attention more efficicently. The equation is defined as:

$$\hat{S} = \text{softmax}\left(\frac{Q\tilde{K}^\top}{\sqrt{d_q}}\right)\left(\text{softmax}\left(\frac{\tilde{Q}\tilde{K}^\top}{\sqrt{d_q}}\right)\right)^+ \text{softmax}\left(\frac{\tilde{Q}K^\top}{\sqrt{d_q}}\right) \tag{5}$$

where Q, K are the full query and key matrices, $\tilde{Q}$, $\tilde{K}$ are the landmark queries and keys, and $(\cdot)^+$ denotes the Moore–Penrose pseudo inverse which is a generalisation of a matrix inverse that can be applied to singular matrices, where a regular inverse does not exist.

The figure below contains the architecture for TransMIL. The idea being that each WSI is cropped into patches and embedded by the features which have been pre-trained by ResNet50.

4.4.7 GAN and MultiPathGAN

Variations in staining protocols and differences in scanner hardware used to digitise WSIs introduce significant heterogeneity in the resulting images. This variability poses a major challenge to machine learning models, often reducing their generalisability and accuracy, as the models struggle to adapt to data distributions not represented in their training sets. To address this issue, Nazki et al. [26] prosed an unsupervised adversarial network called MultiPathGAN, designed to normalise WSIs across diverse data acquisition domains. MultiPathGan preserves the salient structural features of input images while translating their style across domains, addressing inconsistencies in feature representation. A key strength of MultiPathGAN is its ability to train a single generator across multiple domains, improving the robustness and enabling controllable translation towards a target domain. This adaptability is useful in settings where the preservation of histological detail is critical. Figure 11 illustrates the architecture and flow of information of MultiPathGAN.

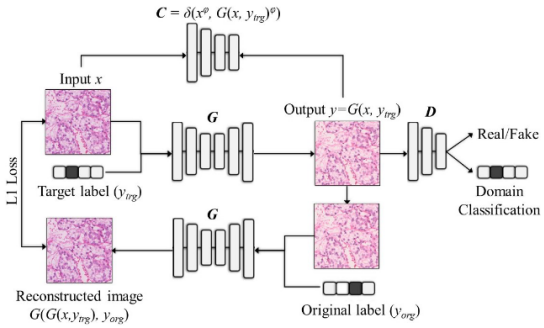


Figure 4: MultiPathGAN and the flow of information [26]

The MultiPathGAN was tested against the data of 120 cases of kidney cancer from the pathology archives in the NHS and it was found that the MultiPathGAN provided a high quality of translation, appropriately adapted the style of images and correctly preserved the anatomical structures. It was found to effectively translate and normalise data from an unseen subset of the WSI space into a known domain without retraining the network. This network was a step in to the right direction in regards to domain translation and addressing issues in data variability.

4.4.8 Self-Supervised Learning

Self Supervised learning (SSL) is a method of learning that does not require labels or annotations on patches. SSL is used for for patch encoding in WSIs.It aims to construct a visual representation using the supervision formulated by the data itself. The learned representations can be used to further improve the performance in various downstream tasks.

4.4.9 Feature Extractors

Recent advancements in self-supervised learning have led to the development of foundational models. These models are trained on unlabelled histopathology datasets. By leveraging contrastive learning, and self supervised learning authors Azizi et al., Chen et al. [27] [28] [29] have introduced these models which can capture morphological features that generalise well and perform well in tasks such as classification, and tumour detection. These models highlight the growing potential of self-supervised foundation models.

4.4.10 Contrastive Learning

Contrastive Learning is a representation learning paradigm that trains an encoder so positives such as two views of the same underlying tissue map to nearby points in embedding space, while negatives map further apart. It is able to avoid label supervision by creating positives via cross-modal pairing. In WSIs, contrastive learning produces embeddings where morphologically similar patches cluster and tissue not similar separate.

H

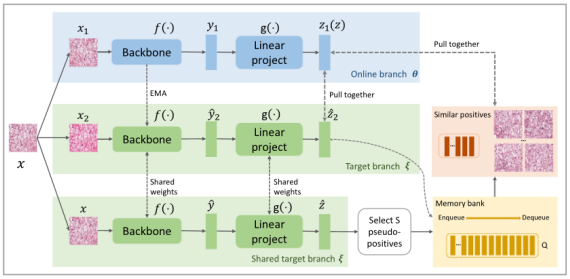


Figure 5: An SRCL approach

4.4.11 CTransPath

4.4.12 Limitation of Traditional Contrastive Learning in Histopathology

There are three major limitations of traditional contrastive learning.

1. Biased Contrastive Pairing and Misidentification of Semantically Correlated Instances: Traditional contrastive learning frameworks treat two augmented views of the same patch as a single positive pair, while other instances are labelled as negatives. This binary-instance level supervision assumes that semantically similar instances exist only within the same image. However, WSIs often contain multiple spatially distinct regions that are morphologically and semantically similar such as patches from the same tissue. These patches may be incorrectly misidentified as negatives. As a result, the learned feature space lacks proper alignment across structurally relevant but spatially separated areas.

2. Local Feature Bias due to CNN- only Architectures Existing contrastive learning methods are built upon Convolutional Neural Networks which emphasis local pattern recognition rather than global. CNNs are suitable for capturing low-level texture features such as cell morphology, and staining granularity they are however, limited by their receptive field. This restricts their ability to capture and interpret global contextual relationships across a WSI. As a result, CNN-based models often fail to represent the hierarchical structure of histopathological images, where both local cellular features and global tissue patterns contribute to diagnostic accuracy.

3. Insufficient Data Variability in SSL pre training Self-supervised pre-training images often suffer from a low data diversity, especially when relying on datasets with unlabelled datasets with sample heterogeneity. A lack of diverse data hinders the ability of the contrastive learning to generalise across datasets.

   These limitations portray the inadequacy of existing SSL strategies to model complex local and global feature dynamics. To address these problems authors Xiyue Wang et al. [30] proposed a transformer-based unsupervised contrastive learning model known as CTranspath. To address, the limitations the semantically-relevatn contrastive learning (SRCL) framework annd a hybrid CNN-transformer backbone was proposed.

The figure above is the architecture of the CTransPath feature extractor. The first part is the CNN structure which employs the Swin Transformer framework, where the patch partition is replaced

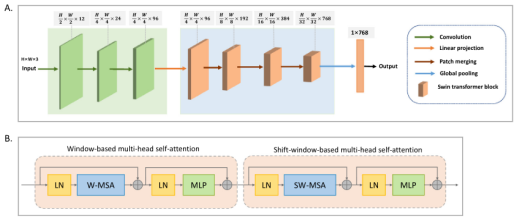


Figure 6: Overall CTranspath feature extractor

by a CNN structure. The CNN part is designed similar to the ResNet Structure, Swim Transformer is designed to generate the hierarchical feature representation using 4 sequential sequences. The second part of the image is the Swin Transformer block, which contains a window-based multi-head self-attention layer and a shift-window-based multi head self-attention (SW-MSA) layer.

4.4.13 Conch and Conch1.5

Proposed in the paper titled "A vision-language foundation model for computational pathology" authors Ming Y. Lu et al introduced a feature extractor which unlike other feature extractors leverages language in addition to extract feature vectors from whole slide images. [31].

CONCH marks a step away from conventional, vision only feature extractors by incorporating natural language supervision alongside image data to learn specific representations.

CONCH is trained in a task agnostic manner on a variety of images and biomedical text using around 1.17 million image caption pairs;.l Its architecture includes an image encoder and a text encoder that project into a shared embedding space, plus a lightweight multimodal decoder for captioning. Training combines a CLIP-style contrastive alignment loss,which pulls matched image text pairs together and pushes mismatched pairs apart, with an caption objective that learns to generate the caption for an image.

A benefit of this is that it is able to classify a tile or slide by comparing it's embedding to a set of class prompts and assigning them with the highest cosine similarity. To reduce sensitive embedding to a set of class prompts and assign the class with the highest cosine similarity. To reduce the sensitivity, paraphrasing prompts were introduced per class. Across a wide range of tasks CONCH returned a strong performance, even outperforming other vision language models such as PLIP, BiomedCLIP, adn OpenAI clip.

4.4.14 GigaPath

Prior models resorted to subsampling a small portion of tiles of each slide, thus missing important slide-level context. GigaPath is a whole slide pathology foundation model pre trained on 1.3 billion 256*256 pathology tiles in 171,189 whole slide images from Providence. Gigapath was proposed to be a model with a massive dataset

Information about the following Foundation Models: A foundation model have been developed for biomedical domains where labelled data is scarce but unlabelled data is in abundant supply. 3 major limitations hindering the development of foundation models for real world applications. Data

availability- pathology data is relatively scarce and vary in quality which limit the performance of foundation models that are pretrained on this data. An example of this is that current existing pathology models are pre trained on WSIs from the The Cancer Genome Atlas (TCGA), a dataset with around 30,000 WSIs or 208 million image tiles on 31 major tissue types. While they are a large resource, TCGA data might not be sufficiently large to address challenges around pathology practices such as heterogeneity and noise artifacts which lead to a significant dip in performance when using TCGA based predictive models and biomarkers. A second problem is the difficulty in designing an architecture that effectively captures both local and global patterns across slides. Current methods treat each slide as an independent pattern and using MIL and thus limiting its ability the ability to model global dependencies. The authors of this foundational model propose GigaPath, an open-weight pathology foundation model which helps alleviates the problems mentioned above. GigaPath uses the image tiles in a pathology slide as input and outputs the slide-level embeddings that can be used as features for diverse applications. Gigapath excels in long-context modelling through distillation of varied local structures and integrating global signatures across the whole slide. GigaPath is made up of a tile encoder to capture the local features and a slide encoder for the global features. The tile encoder is responsible for projecting all the tiles into compact embeddigs. The slide encoder then inputs the sequence of tile embeddings and uses them to generate contextualised embedding taking into account the entire sequence with a transformer. The tile encoder is pre tained on DINOv2 while the slide encoder combines masked auto encoder pre training with LongNet. LongNet is a recently developed method designed for ultra long-sequence modelling.

4.5 Dissertation Formulation

While the literature demonstrates substantial advances in self-supervised learning and foundation models, there remains a gap in systematic benchmarking of these models to determine their suitability for specific downstream tasks. In particular, there have been few studies that rigorously evaluated the performance of modern feature extractors across diverse pathology use cases, limiting the clinical utility. A recent study conducted by Wölflein et al [1] challenged existing assumptions on preprocessing steps such as stain normalisation which had been the norm for digital pathology workflows. With the emergence of feature extractors trained on self-supervised learning this assumption was challenged. Feature extractors have their parameters remain frozen while the aggregator is shallow, but trainable. Previously feature extractors were performed by convolutional neural networks like ResNet-50 however the advancement of self-supervised methods have made training feature extractors without labels a possibility.

An exception to this is the recent work conducted by Wölflein et al. [2], who conducted a comprehensive evaluation of foundation model feature extractors with the attempt to address three fundamental questions:

1. Is Stain normalisation still a necessary preprocessing step?
2. Which feature extractors are best for downstream, slide-level classification?
3. How does magnification affect downstream performance?

To investigate these questions, the authors benchmarked 14 feature extractors across 9 distinct

classification tasks using 5 publicly available datasets. The experiments covered two magnification levels, three downstream MIL-based classification architectures, and a range of preprocessing strategies. Importantly, the study challenged assumptions in the field, regarding the necessity of stain normalisation and the optimal magnification levels for training. Unlike prior benchmarking efforts conducted by (find previous benchmarking attempts if possible), this work place a strong emphasis on clinical relevance by evaluating performance in weakly supervised slide-level biomarker prediction tasks, reflecting real-world diagnostic scenarios.

By minimising reliance on traditional preprocessing steps and highlighting the robustness of certain feature extractors, this study points towards the possibility of streamlined, scalable workflows in computational pathology. This dissertation build on these findings by further evaluating the alignment between feature extractor choice and downstream task performance, particularly in the context of augmentation sensitivity and tumour segmentation.

5 Methodology

The first initial steps was to install Solid Tumour Associative Modelling in Pathology (STAMP) a repository by Kather Labs [32] which integrates multiple feature extractors within the workflow and to set up the directories. After this was complete it was essential to establish ground truths for the models pre-augmentation. To do this STAMP, was step up to test and validate the data.

For image augmentation, the use of the library pyvips. This library was key to convert images to a matrix form for rotation with certain degrees. For the project it was initially decided that the first augmentations applied would be 45 degrees without additional magnification. This library was used as it is effective when augmenting large volume datasets with large memory demands. When writing the code it was decided to write a read me file with the details regarding the augmentation and scale applied to the whole slide image. Included in the script is the code to automatically save the new augmented images in a separate directory while also applying lossless compression to it in order to save space.

Before any augmentation could be done, it was critical to preprocess the dataset and establish the performance metrics pre-augmentation. With this in mind.

Interestingly, upon testing the rotation of 2 Camelyon17 whole slide images both file sizes saw a significant increase in file size from 1.1 gb and 2.1 gb to 13.1 gb and 21.2 gb respectively. Potential reasons for this include: Rotation Introduces Padding or Empty Space

A 60° rotation increases the image canvas size because it has to fit the whole rotated rectangle.

The new image dimensions are larger than the original to accommodate the rotated corners, and the extra areas are filled with background (usually black).

This increases pixel count $\rightarrow$ increases data $\rightarrow$ increases file size. . some things may be causing ineffective compression:

Rotated images are less compressible (LZW works best on patterns, which rotation destroys)

There was a problem encountered with the rotation script using pyvips as the library. This is due

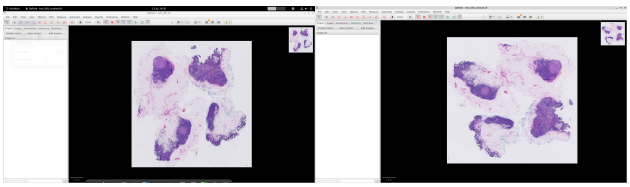


Figure 7: Rotated images 1st test. Original image on the left. Rotated Image is on the right

to the fact that pyvips only supported rotations of 90,180 and 270 degree rotations. This constraint meant that despite the advantages that pyvips brought such as the high data transfer speeds and image readability, it did not support pure rotations as rather than actually rotate the image to any angle it was able to use a workaround where the image was mapped and flipped on the x and y axis as opposed to rotated. This is more efficient with regards to memory managment and therefore quicker to compute. With this information in mind there was another implementation to apply pure rotations was worked on. TO do this the open slide library was used along with the PIL/ Tiff Image Path to process and apply rotations with any angle in mind.

Rotation script was not correct. Currently rewritten the rotation script. There was a memory constraint when loading the wsi images. To overcome this, the WSI images were split into smaller patches and each patch is loaded onto the memory. After this was done the patches were joined together and then the rotation was done with the use of a 2D rotation matrix with the use of sin and cos. Along with this, in order to keep the file size down, lossless compression was used. Here (tiff delfate/lzw/jpeg2k?) was used to achieve this.The rotation script now makes use of a 2d rotation matrix along with a sin and cosine graph to remap the height and width of the image. Using the cv2 library to also make use of warp Affine transformations. The method get rotation matrix finds the centre of the image, and from there applies a rotation based on the angle and scale specified by the user in the command line argument. Once it has been given this information it is able to process the WSI according to what was specified. The rotation script was written to support multiple file types such as the tif, tiff and svs file types. These are the most common file types in relation to whole slide images. Within the script other additions include being able to specify the output directory where the rotated images are to be saved, if this directory does not exist it will create it for the user, it will append the file name to add the word rotated to the original file name. It automatically generates a readme file with the angle that it has rotated the images, the directory it is saved in, the scale of the augmentation as well as how long it took to apply the augmentation.

Software called QuPath was used to examine the results of running the rotation script.

Initially tested on two images from the Camelyon-17 dataset. Currently testing the rotation script. New rotation script need to explain how it works here. This is the output: Version 3 of rotation script (if needed): Fill in if current rotation script does not work or takes to long to run. Upon increasing the zoom.

Further zooming in:

This suggests the tiles were not place in the correct location Attempt: Issue with the reconstruction of the image. Perhaps this is due to the memmap canvas flush with the memory map is causing an issue. Overall shape suggests that it has been rotated. Unclear as to whether the WSI artefacts

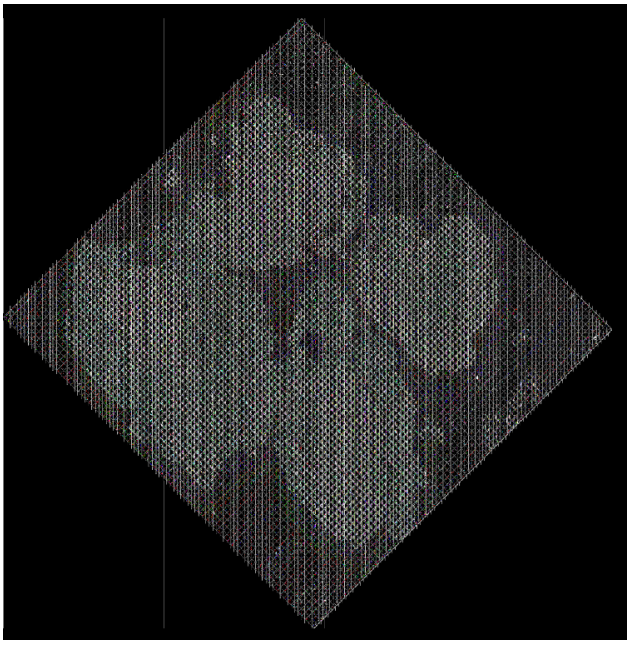


Figure 8: Rotation script attempt

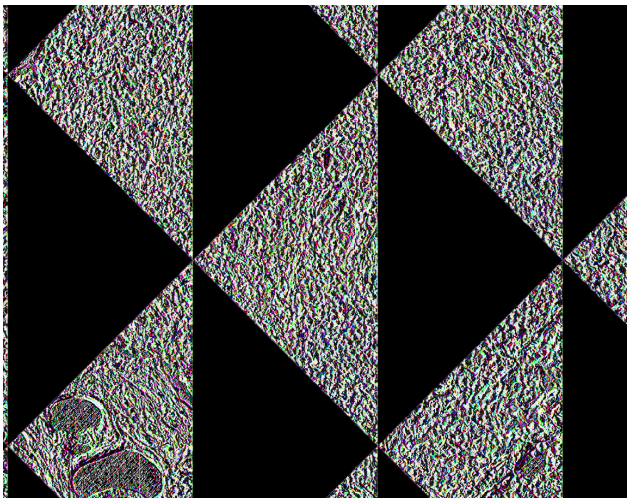


Figure 9: Upon zooming in

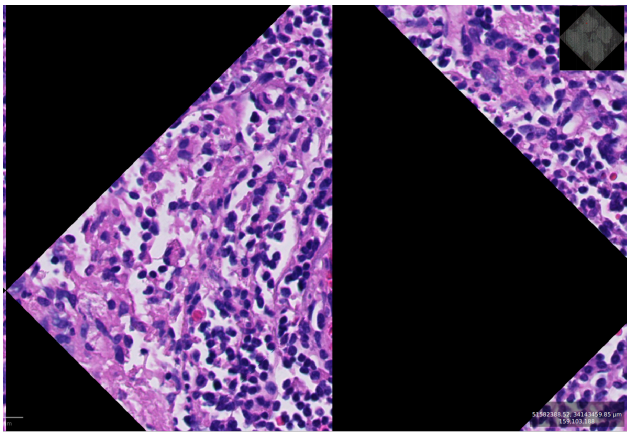


Figure 10: Rotation attempt 3

have been preserved as intended.

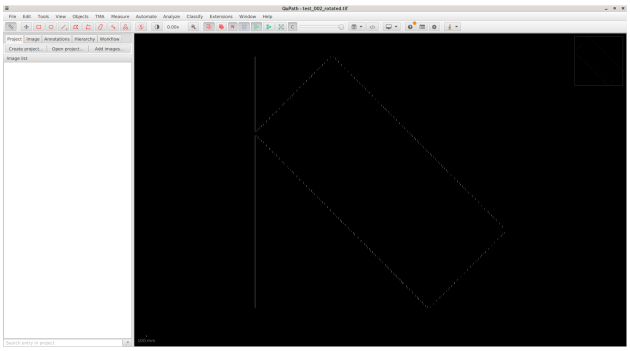


Figure 11: Rotation test 4

5.1 Second approach

The goal of this approach was to replicate the image formatting and storage process used during the digitisation of whole slide images (WSIs). The original images were saved in the ScanScope Virtual Slide format (.svs), a format developed by Aperio (now part of Leica Biosystems). Internally, an .svs file is a specialised form of a TIFF file that supports a pyramidal structure, meaning it contains multiple downsampled resolutions of the same image arranged in levels. This structure enables efficient zooming and panning, which is critical for high-resolution pathology workflows.

Each .svs file typically stores the highest-resolution image (i.e., the full-resolution whole slide) in the base layer (also called IFD 0), with progressively lower-resolution images (e.g., 1/4, 1/16, 1/64 downsampled) stored in SubIFDs. These lower-resolution layers are often compressed using JPEG, while the full-resolution layer may use lossless compression such as LZW, though in some implementations it too may be JPEG-compressed for size efficiency. Additionally, .svs files embed important metadata such as magnification levels, scan dimensions, and acquisition settings.

Understanding this structure was crucial when attempting to apply augmentations (e.g., rotation) to WSIs. Early versions of the augmentation script ignored the pyramidal format and attempted to

rotate and store the full-resolution image using lossless compression alone. This led to significant memory overhead. For instance, an original .svs file of approximately 220 MB ballooned to over 8.5 GB after rotation, due to uncompressed pixel data and lack of efficient tiling or compression. These large files often caused software like QuPath to crash and were unsuitable for practical use.

Moreover, the initial attempts did not preserve the multi-resolution pyramid or the essential TIFF metadata required for correct interpretation by digital pathology tools. Addressing these challenges required designing a rotation pipeline that operated out-of-core (in chunks), padded the rotated regions to avoid edge artifacts, and re-encoded the final result into a pyramidal TIFF structure with appropriate compression and metadata restoration.

This solution did not work correctly as it rotated each chunk to the angle specified rather than the full image. The reconstruction was successful in this regard. Figure 10 contains the new attempt reconstructed correctly. This solution the relational data was modified, while it had correctly reconstructed the image to represent the test image it had not preserved the relational data between each chunk.

The working solution sets out to rotate a whole slide image. The first and essential section of reading a whole slide image is to read the metadata. This metadata is essential to preserving the detail of the WSI and reconstructing the original .svs into a tiff image correctly in the pyramidal structure. The second step is to compute the rotated canvas. Here the code assigns theta which the user specifies in degrees. This is converted to radians and then it builds a rotation matrix as seen below:

$$R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

In order to rotate the four corners of the image around a global centre of rotation. This centre of rotation is to ensure that the previous issue with each chunk being individually rotated does not happen and to preserve relational and spatial information. The algorithm rotates the 4 corners of the image around the global centre (W/2, H/2) to find the bounding box of the rotated image. While computing the bounding box size. It also calculates the offset of the canvas. The trick with this solution was to create a memory mapped array on the disk to avoid writing the WSI to memory. For the reconstruction it was important to divide the slide into several strips or segments to make it manageable on the disk. Then it is iterated over to check the pixel location and ensure there are no gaps in the image reconstruction. Finally the JPEG pyramid is saved with pyvips in order to reconstruct the image into the same file format as the original. After the image has been reconstructed the metadata is restored into the image and the image is reconstructed.

5.2 Working Rotation Solution

The final solution proceeds in four key stages, each critical for accurate, out-of-core rotation and faithful pyramidal reconstruction of a whole-slide image (WSI):

1. Metadata Extraction: The first step it to open the IFD of the input SVS/TIFF and read all of its tags (dimensions, compression, pyramid layout, scanner parameters, etc.). Preserving this metadata is extremely important with , no viewer (QuPath, OpenSlide, etc.) can reassemble tiles into the correct pyramidal hierarchy.

Figure 12: Time taken to rotate 100 WSIs 30 degrees

2. Global Rotation Canvas: The user's desired rotation angle (in degrees) is converted to radians ($\theta = angle\pi/180$), and a standard 2×2 rotation matrix

$$R(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

is built. By rotating the four image corners around the image centre ((W/2, H/2)), we compute the full rotated bounding box (new width/height) and its X/Y offset. This global centre-of-rotation step prevents each tile from drifting independently and maintains correct spatial relationships.

3. Disk-Backed Strip-Wise Warp: A large memmap on disk 'rotated.dat' is allocated at the computed canvas size. We then process the image in overlapping 2000×2000 blocks (adjustable) on the GPU. For each output block, it:

1. Inverse-map its pixel grid back to source coordinates via the inverse rotation matrix.
2. Read only the minimal source patch via OpenSlide.
3. Use PyTorch's 'gridSample' for a single bilinear warp, producing a fully rotated tile.
4. Write that tile directly into the memmap.

This strip-wise approach uses constant, bounded GPU memory and avoids ever holding the entire WSI in RAM.

4. Pyramidal Output and Metadata Reinjection: Finally, we call VIPS's 'tiffsave –pyramid' on the disk memmap to produce a multi-resolution, JPEG-tiled TIFF. We then open that pyramid, write out Though the solution works due to hardware constraints the time taken to rotate the CPTAC-COAD dataset was extremely long around 40 hours Due to hardware constraints the rotation time took a significant amount of time to complete.

5.3 Analysis metrics

Whole-slide images have no canonical orientation, so robustness to flips and rotations is desirable. Wolfein makes the observation in his paper: [2] that "Lunit-DINO excels in term of robustness to right-angle rotations and flips - a much desired property considering that WSIs, unlike natural images, lack a canonical orientation." Researchers Kang and et al. employed these augmentations for this reason, incentivised rotated/flipped embeddings to be close in latent space. Although Lunit Dino is more robust to the horizontal flips rather than vertical ones with non right angles causing the greatest displacement in latent space aside from perspective warp.

Latent space is the vector space of internal representations the encoder maps inputs to. For WSIs, each tile becomes a high-dimensional vector intended to capture the tile's salient histologic content;

tiles that lie semantically close together, while dissimilar tiles fall apart. Due to raw pixel space being huge and redundant, the analysis proceeds with these feature vectors.

The preprocessed represents high dimensional vectors. One effective method of effectively analysing the data is to investigate the dimensionality of the data to assess whether methods such as principal component analysis, t-distributed stochastic neighbour embedding (t-SNE) or uniform Manifold Approximation and projection (UMAP) are best suited to lower the dimensions of the feature vectors and to visualise them.

5.3.1 Uniform Manifold Approximation and Projection for Dimension Reduction:

Proposed by Lealand McInnes et al. in their paper [33], the Uniform Manifold Approximation is a manifold learning technique for dimension reduction. Due to the high dimensionality of the data and some initial exploratory tests were done to assess whether rotation affected the feature extractors at the preprocessing state. To determine whether the 60 rotation had an impact on the preprocessing stage and whether this could lead to a performance drop against the ground truth data it was determined that UMAP was a method that was best suited to model the feature vectors of the extracted tiles. UMAP uses local manifold approximations and patches in union with local fuzzy simplicial set representations to construct a topological representation of the high dimension data. Then with the low dimensional representation of the day, a similar process is used to construct a topological representation. UMAP then optimises the layout of the data, to minimise the cross-entropy between the topological representations.

High-dimensional space. Given points $\{x_i\}_{i=1}^n$ and a distance $d(\cdot,\cdot)$, build a k-NN graph and, for each i, choose ρ_i (local connectivity) and σ_i (local scale) via

$$\rho_i \;=\; \min_{j\in \mathrm{NN}_k(i)} d(x_i,x_j), \qquad \text{and choose } \sigma_i \text{ s.t. } \sum_{j\in \mathrm{NN}_k(i)} \exp\Big(-\frac{\max\{0,\; d(x_i,x_j)-\rho_i\}}{\sigma_i}\Big) \;=\; \log_2 k.$$

Define the (directed) membership strengths

$$\mu_{i\to j} \;=\; \exp\Big(-\frac{\max\{0,\; d(x_i,x_j)-\rho_i\}}{\sigma_i}\Big).$$

Symmetrise with the fuzzy union to obtain undirected edge memberships

$$p_{ij} \;=\; \mu_{i\to j} + \mu_{j\to i} - \mu_{i\to j}\,\mu_{j\to i} \quad \in [0,1].$$

Low-dimensional space. Place $\{y_i\}_{i=1}^n \subset \mathbb{R}^2$ and measure low-D affinities with

$$q_{ij} \;=\; \frac{1}{1 + a\,\|y_i - y_j\|^{2b}},$$

where $a, b > 0$ are chosen (numerically) so that q_{ij} approximates a flat regime up to `min_dist` and an exponential tail thereafter (controlled by `spread`).

Objective (cross-entropy between fuzzy sets). UMAP finds $Y = \{y_i\}$ by minimising

$$\mathcal{L}(Y) \;=\; \sum_{i<j} \Big(-\; p_{ij}\; \log q_{ij} \;-\; (1 - p_{ij})\; \log\big(1 - q_{ij}\big) \Big).$$

Optimisation. In practice, $\mathcal{L}$ is optimised by stochastic gradient descent with positive edges sampled proportional to p_{ij} and negative edges sampled uniformly, yielding attractive forces for large p_{ij} and repulsion otherwise.

Notes. $k =$ `n_neighbours` controls locality; `min_dist` tunes cluster tightness via (a, b); the `metric` (e.g., cosine or euclidean) sets the distance function $d(\cdot, \cdot)$.

5.4 T- distributed Stochastic Neighbour Embedding (t-SNE)

T-distributed Stochastic Neighbour Embedding is another key dimensionality reduction metric proposed by Laurens van der Maaten and Geoffrey Hinton [34] used in this dissertation to visualise the geometry of tile embeddings produced by the feature extractos in the STAMP pipeline.t-SNE converts pairwise distance in the original feature space into probabilities that encode neighbourhood relationships, then find a 2D map whose pairwise probabilities match the original ones as closely as possible. Due to it strongly preserving local neighbourhoods, it is well suited to testing whether the rotated tiles land in the same regions of representation space as the pre-rotated counterparts.

t-SNE Formulation

Given a dataset of N high-dimensional points $\{x_1, x_2, \ldots, x_N\}$, t-SNE seeks to find a low-dimensional representation $\{y_1, y_2, \ldots, y_N\}$ with $y_i \in \mathbb{R}^d$ (usually $d = 2$ or 3).

High-dimensional similarities

For each pair of points (x_i, x_j), define the conditional probability that x_i would choose x_j as its neighbour:

$$p_{j|i} = \frac{\exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(-\frac{\|x_i - x_k\|^2}{2\sigma_i^2}\right)},$$

where σ_i is chosen such that the perplexity of the distribution matches a predefined value:

$$\text{Perp}(P_i) = 2^{-\sum_j p_{j|i} \log_2 p_{j|i}}.$$

The joint probability in the high-dimensional space is symmetrised as:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}.$$

Low-dimensional similarities

For the corresponding low-dimensional points (y_i, y_j), define the similarity using a Student-t distribution with one degree of freedom:

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}}.$$

Objective function

The aim of t-SNE is to make the distribution $Q = \{q_{ij}\}$ match $P = \{p_{ij}\}$ as closely as possible. This is done by minimising the Kullback–Leibler(KL) divergence:

$$C = \text{KL}(P\|Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

Optimisation

The embedding points $\{y_i\}$ are optimised via gradient descent in order to minimise C.

6 Results

When comparing models built from different feature extractors, evaluation should be performed at a slide level rather than the patch level. This is due to tiles being sampled from the same WSI and therefore the same patient. This makes them highly correlated and inherits the same label. Treating patches as independent observations can lead to an inflation in the sample size of positives, which attributes to information leakage between training and test splits and can lead to higher observed positive rates. Slide level classification obtained by aggregating patches via MIL or other methods mentioned above better reflect the clinical decision unit and yield more reliable estimates of generalisation.

All the analysis in this work use slide grouping. In the CPTAC-COAD dataset used in this project, there are slides from 103 patients with evaluation done on 5 fold cross validation.

6.0.1 Classification Metrics

There are two primary metrics used in this analysis Area Under the Receiver Operating Characteristic curve (AUROC) The AUROC summaries how well the model ranks mutant or positive slides above wild-type or negative slides across the the classification thresholds. The area under the curve represents the true positive rate against the false positive rate. Probabilistically the area under the curve equals the likelihood that a randomly chosen positive slide receives a higher score than a randomly chosen negative one. In imbalanced settings where a tissue can be seen to be rarer than the other the AUROC can be more positive because the false positive rate averages over a larger number of negatives leading to inaccuracies.

The second metric is the area under the precision recall curve. AUPRC focuses on the positive class by integrating precision, a positive predictive value measure the true positive of the slides that

are classed as positive against the recall, which represents the number of true positive slides that have been correctly identified. In settings where positives are rare the PRC is a more informative measure.

6.1 Pre Augmentation Data against Ground Truth

6.1.1 Conch1.5

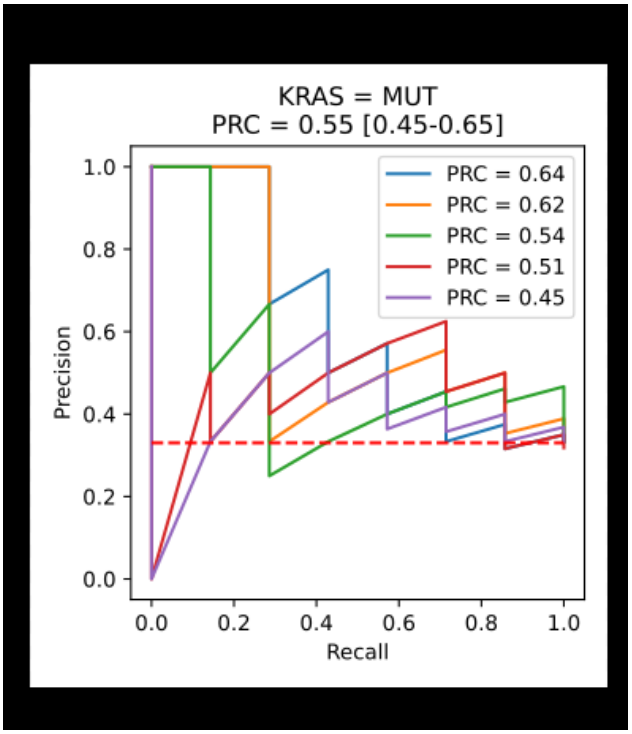


Figure 13: PRC CONCH Rotation

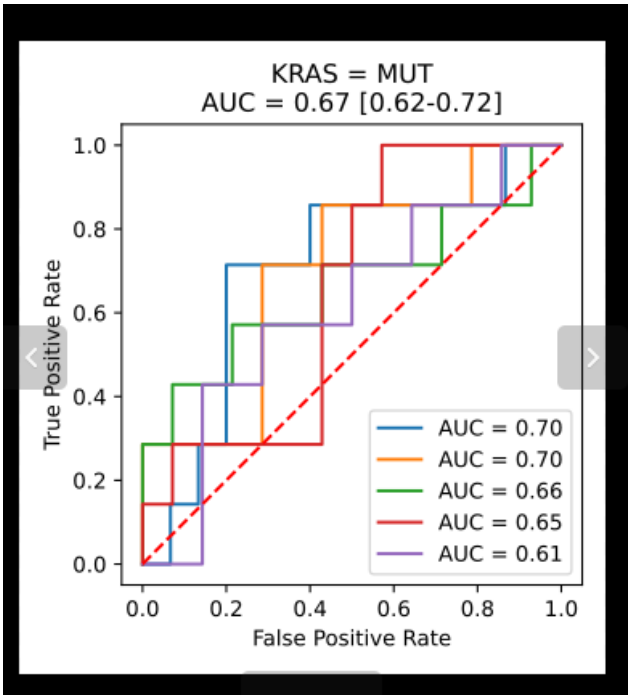


Figure 14: AUC CONCH1.5

6.1.2 CTranspath

The figure below contains an example for the preprocessed image. The red around the picture indicates that the preprocessor has removed so as to not use additional resources on redundant space. The image is from the stamp prepossessing as it indicated where the slide is preprocessed from.

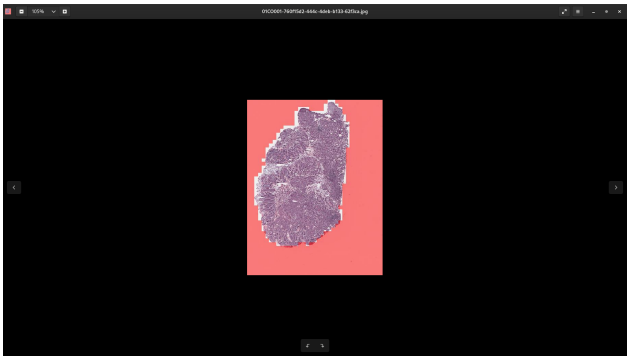


Figure 15: Pre-Augmentation

Figure 14 shows the PRC curve with a mean of 0.59 with a large range over the 5 folds 0.3. The worst performing curve being the purple curve with a low score of 0.39 while the best score is 0.70. This wide range suggests that the model does not generalise well as performance is not consistent. The overall average suggests that the curve has a moderate to low recall but trends toward a more the baseline as the recall increases.

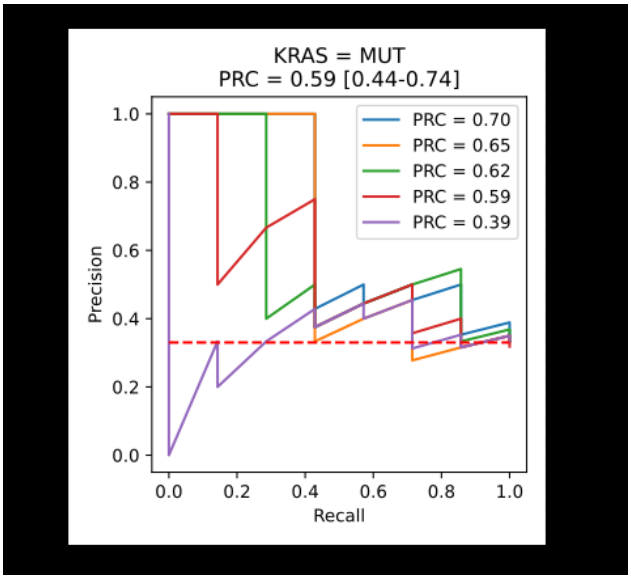


Figure 16: PRC CTranspath

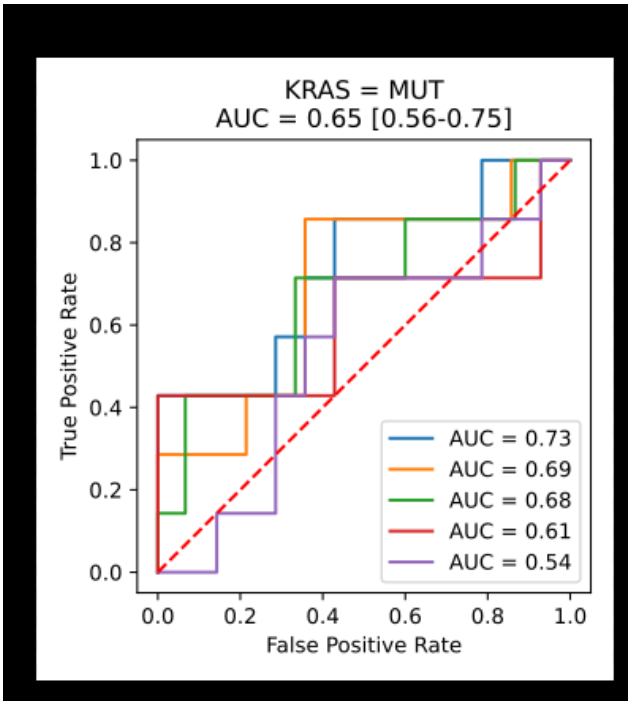


Figure 17: AUC CTranspath

The figure above is the AUC graph.The model performs fairly well on the AUC while suggesting that there is a chance that the model gives the mutant a higher score due to the score being above 0.65

6.1.3 H-Optimus1

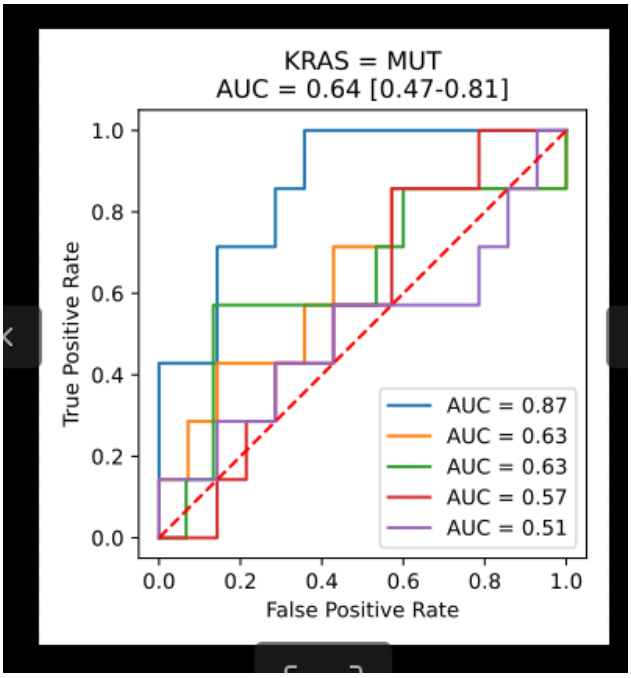


Figure 18: H-Optimus1

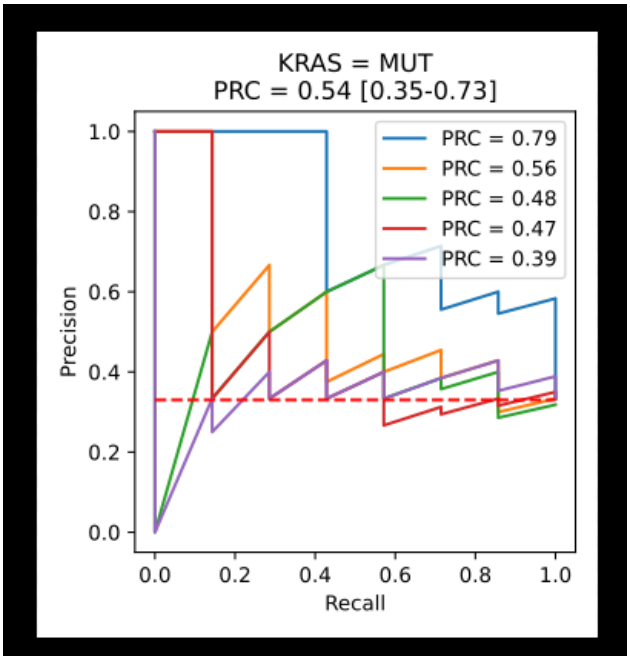


Figure 19: H-Optimus1

6.1.4 Gigapath

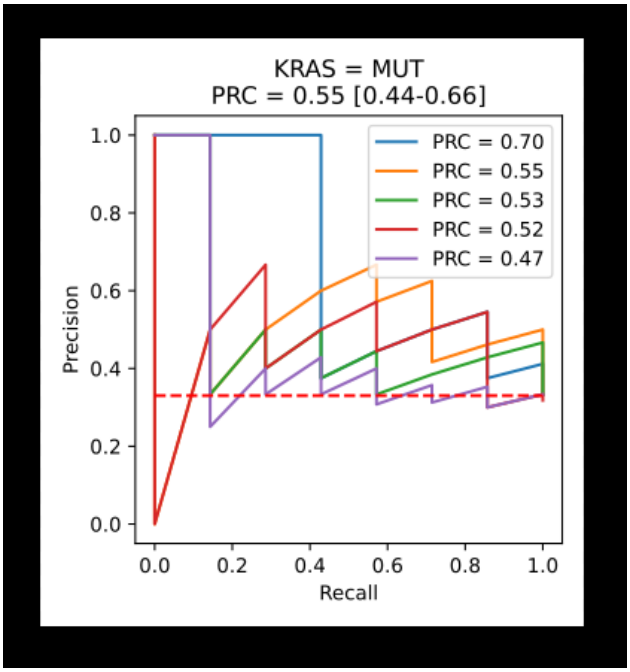


Figure 20: PRC GigaPath

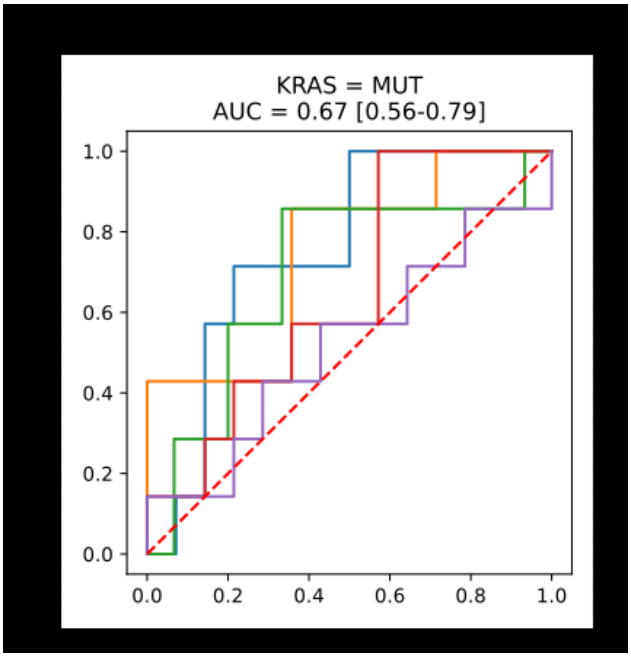


Figure 21: AUC GigaPath

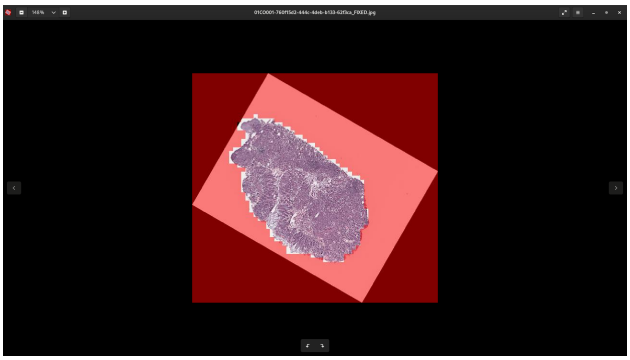


Figure 22: Augmented Preprocessed

6.2 60 Degree Rotation

Figure 22 displays an example output of a preprocessed image. This reflects the workflow of the feature extractor as it sets out to remove the white space by filling it in with red as seen in the figure below. This is so the empty space is ignored and the feature extractors can correctly preprocess the image while ignoring and avoiding redundant space that adds a computational cost to the workflow.

6.2.1 CONCH1.5

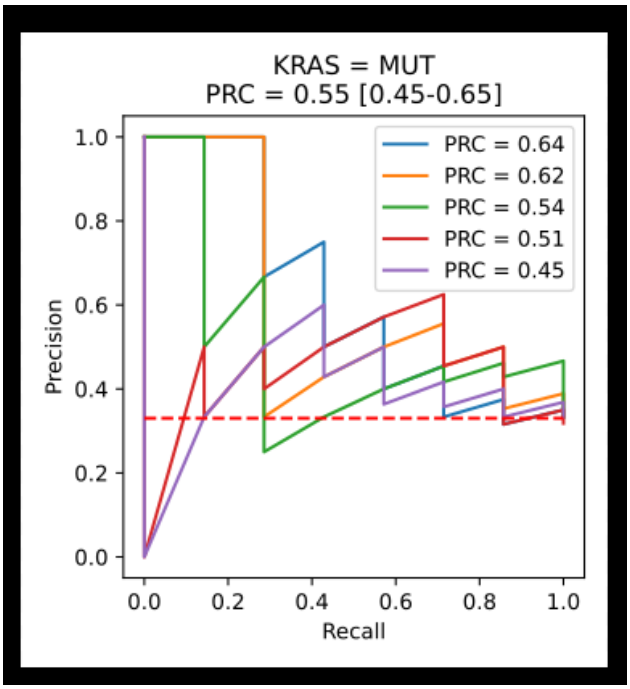


Figure 23: Precision-Recall Data

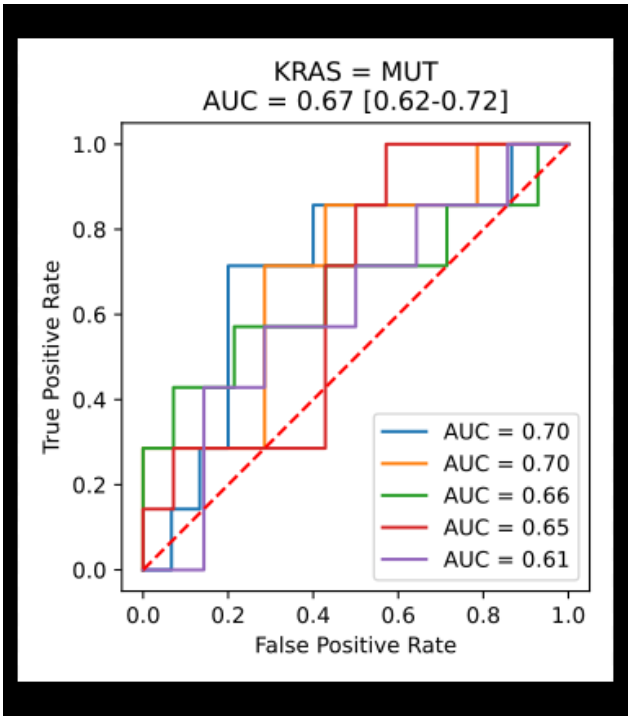


Figure 24: Area Under Curve

6.2.2 Feature Representation

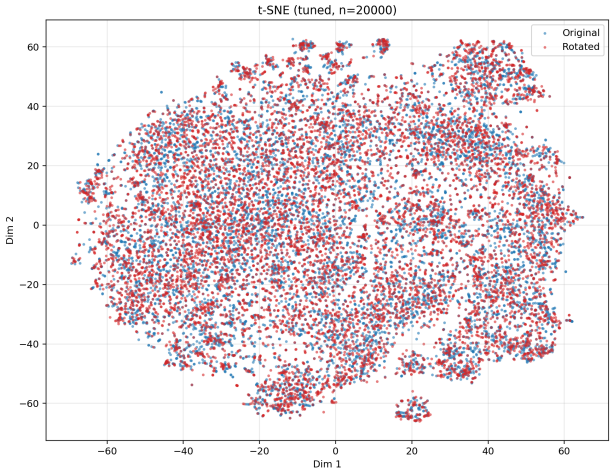


Figure 25: t-SNE Conch1.5

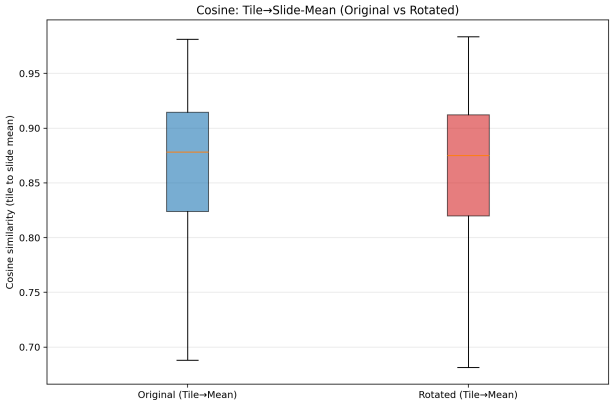


Figure 27: Cosine box plot

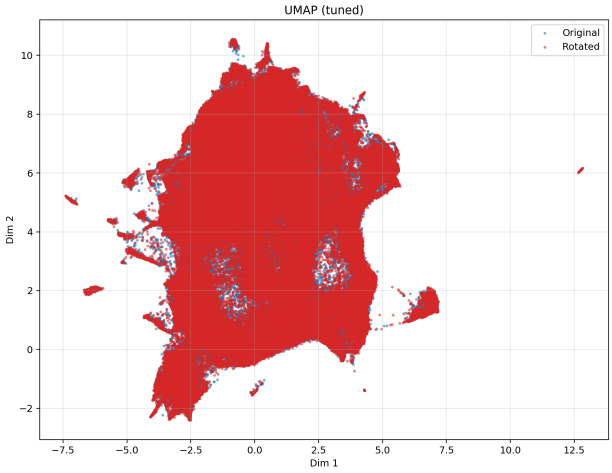


Figure 26: UMAP Conch1.5

6.2.3 CTranspath

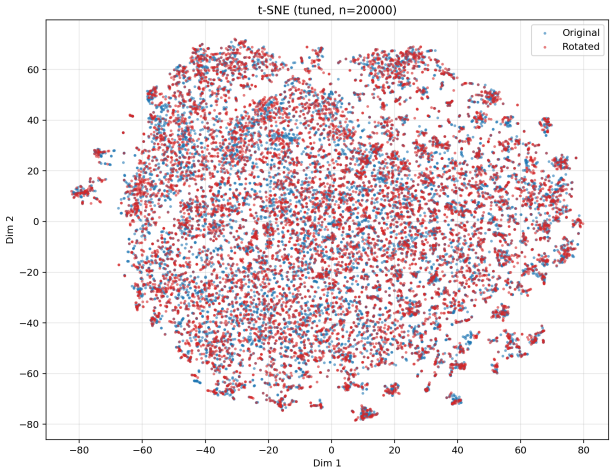


Figure 28: T-SNE CTranpath

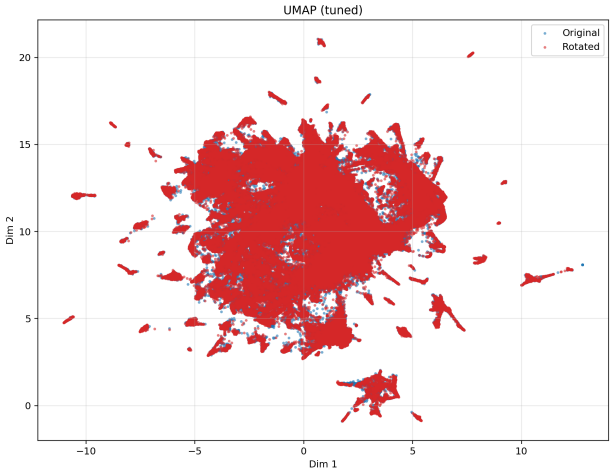


Figure 29: UMAP CTransPath

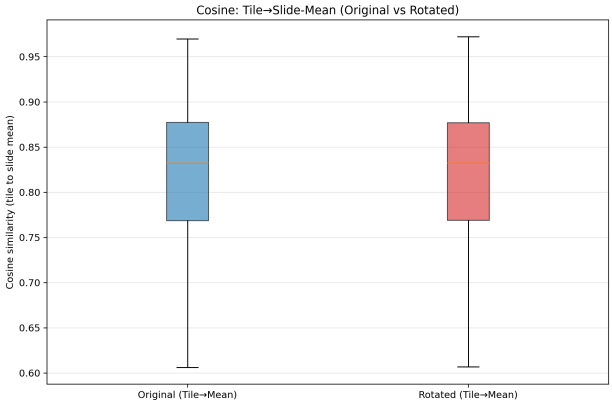


Figure 30: Cosine Boxplot

6.2.4 GIGAPATH

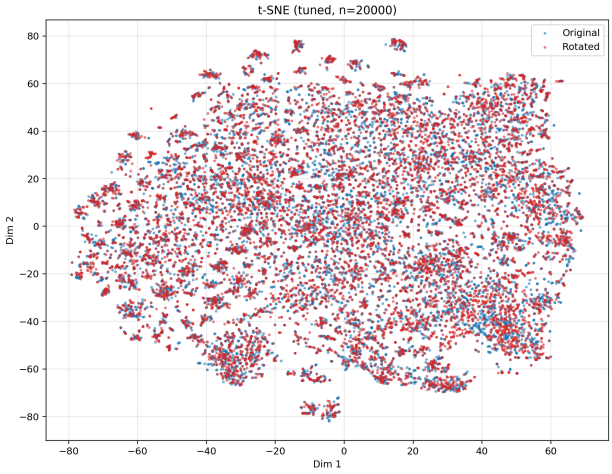


Figure 31: t-SNE tuned

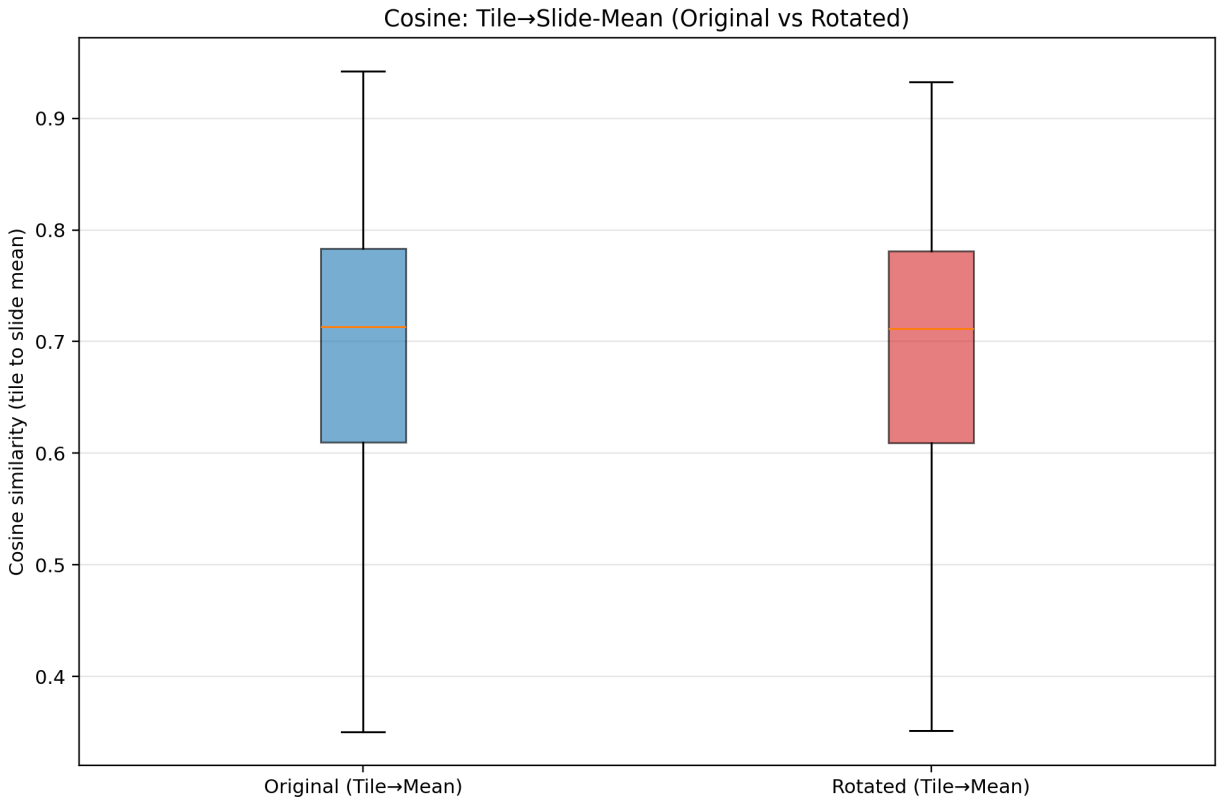


Figure 33: Cosine box plot

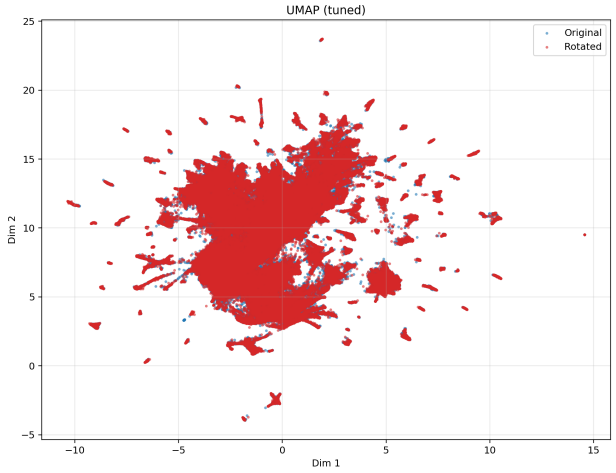


Figure 32: Umap Gigapth

After running UMAP and TSNE that there is a large overlap of both rotated and non-rotated images at 60 degrees. This suggests that the rotation has very little effect on the representation.

6.2.5 H-Optimus-1

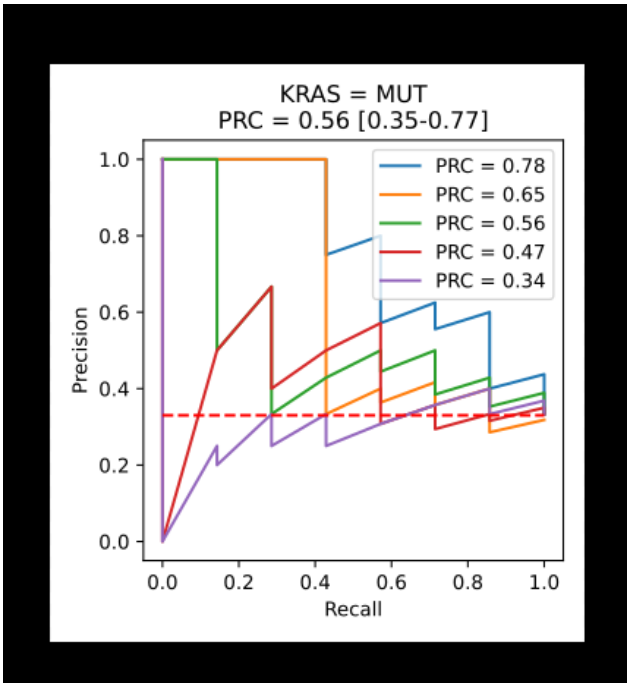


Figure 34: H-Optimus Rotated

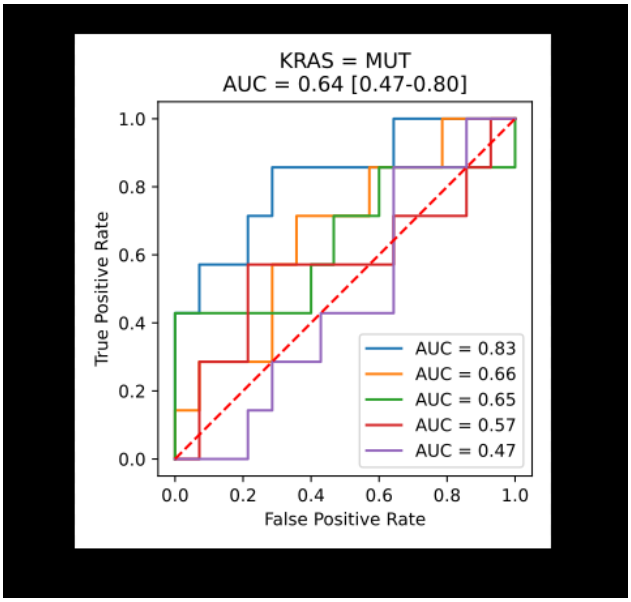


Figure 35: H-optimus AUC

7 Evaluation

Overall, the instability in the data suggests the dataset was too small hence the curve. Based on the t-SNE, UMAP and Cosine similarities between feature extractors it can be concluded that the rotation of 60 degrees has a minimal impact on the patch embeddings on several of the feature extractors. With both the neighbours and the minimum distance of the t-sne and the UMAP largely overlapping, there is no quantifiable impact rotation has on the extraction of feature vectors. Another piece of evidence that supports this claim is the similarity of the cosine similarity. The box plots for each extractor show a slight offset to the median for the original against the rotated.

Across the original tests, Conch 1.5 achieved ROC-AUC of 0.67, 95 percent CI between 0.62 and 0.72 and a PR AUC between 0.45 and 0.6, CTransPath reached of 0.65 between 0.56 - 0.75 and PR AUC of 0.59, between 0.44 and 0.74 and H-Optimus-1 have achieved a high 0.64 (0.47-0.81) and PR-AUC of 0.54 (0.35-0.73). Under a 60 degree rotation the central metrics and intervals fluctuated slightly but remained largely the same. Differences emerge in the stability: H-Optimus exhibits the widest per fold sprread with an ROC-AUC of 0.47-0.83 and a PR-AUC of 0.34 - 0.78 which suggests a sensitivity to data splits. The stair case appearance on the ROC plots is likely due to a small evaluation set whereas CTranspath is the most consisten above the prevision baseline and CONCH1.5 is similar with a little more vairability. Based on this limited evidence it would not be enough to conclude that the feature extractors are entirely robut to rotations.

8 Future Work

Preliminary conclusion: The analysis suggests that a 60 degree rotation has little effect on the tested feature extractors. However, this evidence is insufficient to claim that rotation has no impact on the learned representations. In order to make a more definitive case for that statement, more rigours tests such as comparing the feature distributions using a divergence measure such as KL or an intraclass correlation coefficient. Additional angles such as 30, 75, 135 would increase the validity of the tests. In addtion to more angles , additional feature extractors such as Luinit Dino, Uni and Virchow and even CNN based image extractors such as Res-Net should be tested.

A more comprehensive study would include larger cohorts and external datasets such as the TCGA-BRCA and the TCGA-COAD, to ensure a more robust assessment process. Additionally, the Camelyon-17 dataset could be used to externally validate the rotation robustness.

To further asses the impact of rotation a rigorous test would be to train a model on non rotated image and then to test on rotated images to assess the performance. Another avenue to explore would be train a model on non rotated images and then to test against an unseen dataset with several unseen angles mixed in.

Practical constraints: Due to compute limits and interrupted runs, the 30 degree rotation did not complete fully. Only 275 slides were processed this hindered the capability of the evaluation resources were limited. Another practical constraint was the rotation taking around 30 hours to complete the COAD dataset. In future this would have to be optimised.

9 Conclusion

In conclusion, this dissertation set out to investigate how global rotation affects feature extractors in whole slide images. While the study did not exhaustively resolve this question across the desired angles set out at the begging of the dissertation period , it delivered a workflow capable of rotating whole slide images, standardised feature extraction across multiple levels, slide level aggregation. Using multiple metrics such as UMAP, t-SNE and Cosine similarity diagnostics. Using these tools, some results were obtained to begin to answer the questions set out at the beginning of the dissertation period. Due to certain limiting factors the project could not progress further and provide a clearer answer as to whether or not rotation has a signifiant factor on the performance of feature extractors.

10 Acknowledgements

Wow here we are again. Let me try and stumble through this. Firstly I would like to extend my sincerest thanks to my supervisor Oggie without whom I would not have even dreamed about attempting a project in this field and for giving support and guidance throughout the year and for an opportunity to reach for goals. I would like to thank my family. Dad,Mum, Rayna and Rishi without your endless support and guidance I would not have even have made it to the University. I'd like to thank my friends who listened to me ranting about the problems I was encountering or alongside me on the countless sleepless nights. I'd like to also say that the first time was so nice I had to do it twice. Last but not least, I'd like to thank me, I'd like to thank me for putting in all this hard work. I'd like to thank me for believing in me.

11 Appendix

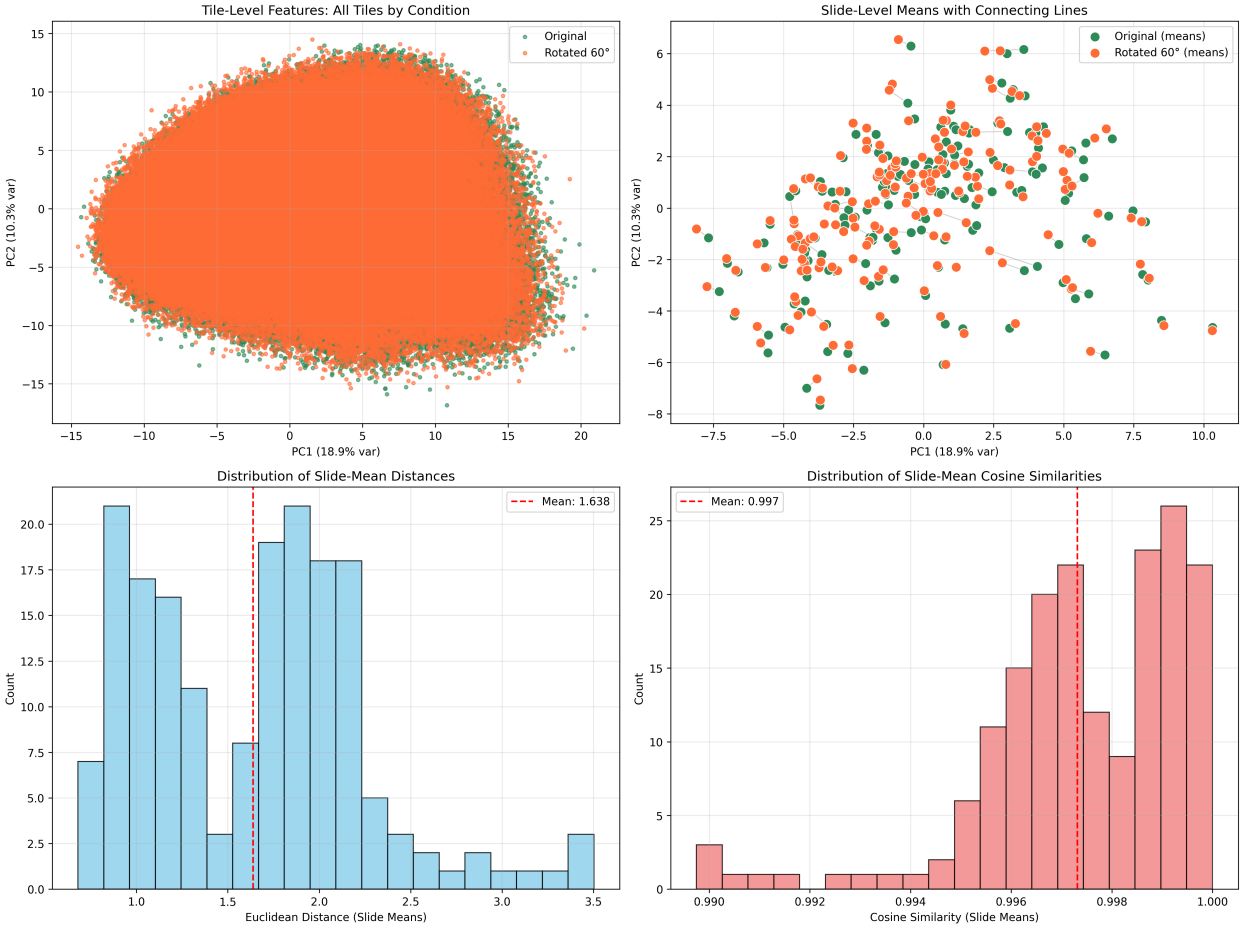


Figure 36: Metrics for Conch 1.5

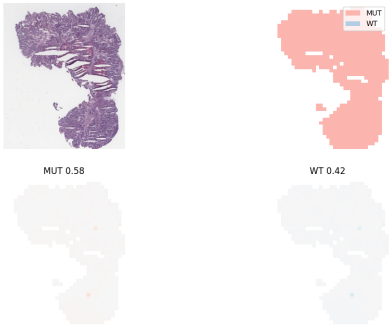


Figure 37: Heatmap of an example image pre-rotation

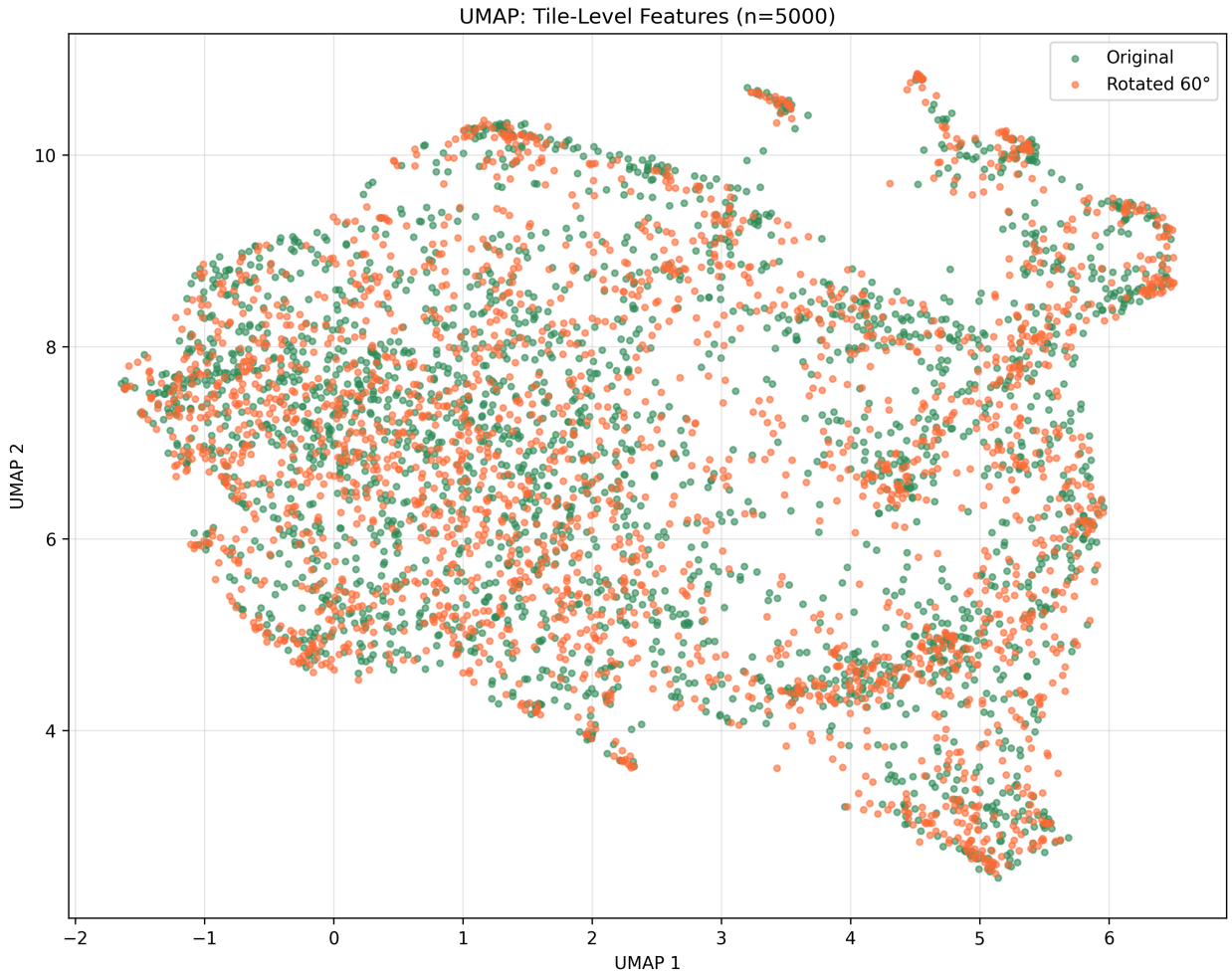
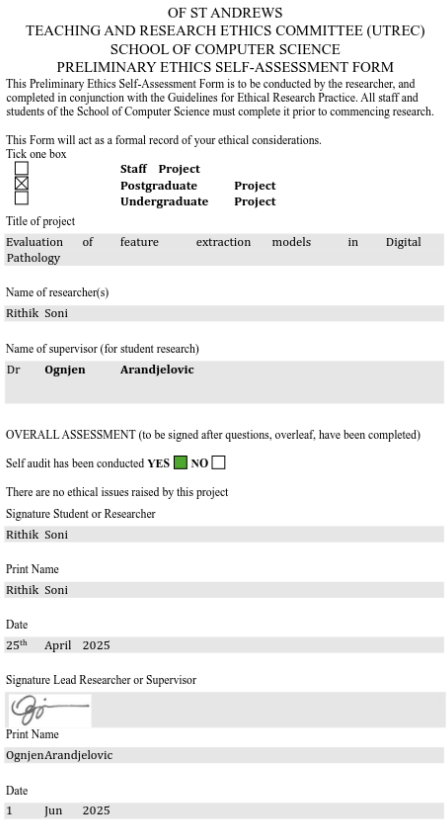


Figure 38: UMAP of CONCH1.5

11.1 Ethics



OF ST ANDREWS
TEACHING AND RESEARCH ETHICS COMMITTEE (UTREC)
SCHOOL OF COMPUTER SCIENCE
PRELIMINARY ETHICS SELF-ASSESSMENT FORM

This Preliminary Ethics Self-Assessment Form is to be conducted by the researcher, and completed in conjunction with the Guidelines for Ethical Research Practice. All staff and students of the School of Computer Science must complete it prior to commencing research.

This Form will act as a formal record of your ethical considerations.
Tick one box

☐ **Staff Project**
☒ **Postgraduate Project**
☐ **Undergraduate Project**

Title of project

Evaluation of feature extraction models in Digital Pathology

Name of researcher(s)

Rithik Soni

Name of supervisor (for student research)

Dr **Ognjen Arandjelovic**

OVERALL ASSESSMENT (to be signed after questions, overleaf, have been completed)

Self audit has been conducted **YES** ☒ **NO** ☐

There are no ethical issues raised by this project

Signature Student or Researcher

Rithik Soni

Print Name

Rithik Soni

Date

25th April 2025

Signature Lead Researcher or Supervisor

Print Name

OgnjenArandjelovic

Date

1 Jun 2025

Figure 39: Ethics Approval

References

[1] G. Wolflein and D. Fenzl, "A good feature extractor is all you need for weakly supervised learning in histopathology." arXiv preprint arXiv:2303.17615, 2023.

[2] G. Wölflein, D. Ferber, A. R. Meneghetti, O. S. M. E. Nahhas, D. Truhn, Z. I. Carrero, D. J. Harrison, O. Arandjelović, and J. N. Kather, "Benchmarking pathology feature extractors for whole slide image classification," 2024.

[3] W. Lingle, B. J. Erickson, M. L. Zuley, R. Jarosz, E. Bonaccio, J. Filippini, J. M. Net, L. Levi, E. A. Morris, G. G. Figler, P. Elnajjar, S. Kirk, Y. Lee, M. Giger, and N. Gruszauskas, "The cancer genome atlas breast invasive carcinoma collection (tcga-brca) (version 3)." `https://doi.org/10.7937/K9/TCIA.2016.AB2NAZRP`, 2016. Data set from The Cancer Imaging Archive.

[4] G. Litjens and P. Bandi, "1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset," *GigaScience*, vol. 7, pp. 1–8, 2018.

[5] I. R. Kong and O. Fenton, "What is precision medicine?," *European Respiratory Journal*, pp. 2–12, 2017.

[6] J. Pallua, A. Brunner, B. Zelger, M. Schirmer, and J. Haybaeck, "The future of pathology is digital," *Pathology - Research and Practice*, vol. 216, no. 9, p. 153040, 2020.

[7] J. F. Boan Lai, "Artificial intelligence in cancer pathology: Challenge to meet increasing demands of precision medicine," *International Journal of Oncology*, vol. 63, no. 107, pp. 1–30, 2023.

[8] U. Catalyurek, M. Beynon, C. Chang, T. Kurc, A. Sussman, and J. Saltz, "The virtual microscope," *IEEE Transactions on Information Technology in Biomedicine*, vol. 7, no. 4, pp. 230–248, 2003.

[9] L. Pantanowitz, A. Sharma, A. B. Carter, T. Kurc, A. Sussman, and J. Saltz, "Twenty years of digital pathology: An overview of the road travelled, what is on the horizon, and the emergence of vendor-neutral archives," *Journal of Pathology Informatics*, vol. 9, no. 1, p. 40, 2018.

[10] V. L. Schumacher, F. Aeffner, E. Barale-Thomas, C. Botteron, J. Carter, L. Elies, J. A. Engelhardt, P. Fant, T. Forest, P. Hall, D. Hildebrand, R. Klopfleisch, T. Lucotte, H. Marxfeld, L. Mckinney, P. Moulin, E. Neyens, X. Palazzi, A. Piton, E. Riccardi, D. R. Roth, S. Rousselle, J. D. Vidal, and B. Williams, "The application, challenges, and advancement toward regulatory acceptance of digital toxicologic pathology: Results of the 7th estp international expert workshop (september 20-21, 2019)," *Toxicologic Pathology*, vol. 49, no. 4, pp. 720–737, 2021. PMID: 33297858.

[11] P. D. Caie and N. Dimitriou, "Precision medicine in digital pathology via image analysis and machine learning," in *Artificial Intelligence in Pathology (Second Edition)* (C. Chauhan, ed.), pp. 233–257, Elsevier, 2025.

[12] E. Jenkinson and O. Arandjelović, "Whole slide image understanding in pathology: What is the salient scale of analysis?," *BioMedInformatics*, vol. 4, no. 1, pp. 489–518, 2024.

[13] G. Campanella and V. Wiesmann, "Terabyte-scale deep multiple instance learning for classification and localisation in pathology." arXiv preprint arXiv:1805.06983, 2018.

[14] E. Jain, A. Patel, A. V. Parwani, S. Shafi, Z. Brar, S. Sharma, and S. K. Mohanty, "Whole slide imaging technology and its applications: Current and emerging perspectives," *International Journal of Surgical Pathology*, vol. 32, no. 3, pp. 433–448, 2024. PMID: 37437093.

[15] T. L. Sellaro, R. Filkins, C. Hoffman, J. L. Fine, J. Ho, A. V. Parwani, L. Pantanowitz, and M. Montalto, "Relationship between magnification and resolution in digital pathology systems," *Journal of Pathology Informatics*, vol. 4, p. 21, 2013.

[16] D. F. Steiner, P.-H. C. Chen, and C. H. Mermel, "Closing the translation gap: Ai applications in digital pathology," *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, vol. 1875, no. 1, p. 188452, 2021.

[17] Y. Xu, J. Zhang, E. I.-C. Chang, M. Lai, and Z. Tu, "Context-constrained multiple instance learning for histopathology image segmentation," in *Medical Image Computing and*

Computer-Assisted Intervention – MICCAI 2012, vol. 7510 of *Lecture Notes in Computer Science*, pp. 623–630, Springer, 2012.

[18] X. Wang, H. Chen, C. Gan, H. Lin, Q. Dou, E. Tsougenis, Q. Huang, M. Cai, and P.-A. Heng, "Weakly supervised deep learning for whole slide lung cancer image analysis," *IEEE Transactions on Cybernetics*, vol. 50, no. 9, pp. 3950–3962, 2020.

[19] M. Mohammadi, J. Cooper, O. Arandelović, C. Fell, D. Morrison, S. Syed, P. Konanahalli, S. Bell, G. Bryson, D. J. Harrison, and D. Harris-Birtill, "Weakly supervised learning and interpretability for endometrial whole slide image diagnosis," *Experimental Biology and Medicine (Maywood, N.J.)*, vol. 247, no. 22, pp. 2025–2037, 2022.

[20] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognition*, vol. 77, p. 329–353, May 2018.

[21] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proceedings of the 35th International Conference on Machine Learning* (J. Dy and A. Krause, eds.), vol. 80 of *Proceedings of Machine Learning Research*, pp. 2127–2136, PMLR, 10–15 Jul 2018.

[22] B. Li, Y. Li, and K. W. Eliceiri, "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14318–14328, June 2021.

[23] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, and y. zhang, "Transmil: Transformer based correlated multiple instance learning for whole slide image classification," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 2136–2147, Curran Associates, Inc., 2021.

[24] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *CoRR*, vol. abs/2005.12872, 2020.

[25] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *CoRR*, vol. abs/2102.04306, 2021.

[26] H. Nazki and O. Anand, "Multipathgan: Structure preserving stain normalisation using unsupervised multi-domain adversarial network with perception loss." arXiv preprint arXiv:2205.00412, 2022.

[27] S. Azizi, L. Culp, J. Freyberg, B. Mustafa, S. Baur, S. Kornblith, T. Chen, N. Tomasev, J. Mitrović, P. Strachan, S. S. Mahdavi, E. Wulczyn, B. Babenko, M. Walker, A. Loh, P.-H. C. Chen, Y. Liu, P. Bavishi, S. M. McKinney, J. Winkens, A. G. Roy, Z. Beaver, F. Ryan, J. Krogue, M. Etemadi, U. Telang, Y. Liu, L. Peng, G. S. Corrado, D. R. Webster, D. Fleet, G. Hinton, N. Houlsby, A. Karthikesalingam, M. Norouzi, and V. Natarajan, "Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging," *Nature Biomedical Engineering*, vol. 7, no. 6, pp. 756–779, 2023.

[28] R. J. Chen, T. Ding, M. Y. Lu, D. F. K. Williamson, G. Jaume, A. H. Song, B. Chen, A. Zhang, D. Shao, M. Shaban, M. Williams, L. Oldenburg, L. L. Weishaupt, J. J. Wang, A. Vaidya, L. P. Le, G. Gerber, S. Sahai, W. Williams, and F. Mahmood, "Towards a general-purpose foundation model for computational pathology," *Nature Medicine*, vol. 30, no. 3, pp. 850–862, 2024.

[29] R. J. Chen, C. Chen, Y. Li, T. Y. Chen, A. D. Trister, R. G. Krishnan, and F. Mahmood, "Scaling vision transformers to gigapixel images via hierarchical self-supervised learning," 2022.

[30] X. Wang, S. Yang, J. Zhang, M. Wang, J. Zhang, W. Yang, J. Huang, and X. Han, "Transformer-based unsupervised contrastive learning for histopathological image classification," *Medical Image Analysis*, vol. 81, p. 102559, 2022.

[31] M. Y. Lu, B. Chen, D. F. Williamson, R. J. Chen, I. Liang, T. Ding, G. Jaume, I. Odintsov, L. P. Le, G. Gerber, *et al.*, "A visual-language foundation model for computational pathology," *Nature Medicine*, vol. 30, p. 863–874, 2024.

[32] O. S. M. El Nahhas, M. van Treeck, G. Wölflein, M. Unger, M. Ligero, T. Lenz, S. J. Wagner, K. J. Hewitt, F. Khader, S. Foersch, D. Truhn, and J. N. Kather, "From whole-slide image to biomarker prediction: end-to-end weakly supervised deep learning in computational pathology," *Nature Protocols*, Sep 2024.

[33] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 2020.

[34] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

[35] B. A. Alexi, "Supporting data for "1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset"." Data set, 2018. `https://doi.org/10.5524/100439`.

[36] N. Dimitriou and O. Anand, "Magnifying networks for histopathological images with billions of pixels," *Diagnostics*, 2024. `https://doi.org/10.3390/diagnostics14050524`.

[37] O. Arandjelovic and N. Dimitriou, "Deep learning for whole slide image analysis: An overview," *Frontiers in Medicine*, vol. 6, no. November, pp. 1–7, 2019.

[38] N. Dimitriou, O. Arandjelović, and D. J. Harrison, "Magnifying networks for histopathological images with billions of pixels," *Diagnostics*, vol. 14, no. 5, 2024.

[39] G. Wölflein, D. Ferber, D. Truhn, O. Arandjelović, and J. N. Kather, "Llm agents making agent tools," 2025.

[40] H. Nazki, O. Arandjelovic, I. H. Um, and D. Harrison, "Multipathgan: Structure preserving stain normalization using unsupervised multi-domain adversarial network with perception loss," in *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, SAC '23, (New York, NY, USA), p. 1197–1204, Association for Computing Machinery, 2023.

[41] R. S. Weinstein, "Prospects for telepathology," *Human Pathology*, vol. 17, no. 5, pp. 433–434, 1986.

[42] N. Kumar, R. Gupta, and S. Gupta, "Whole slide imaging (wsi) in pathology: Current perspectives and future directions," *Journal of Digital Imaging*, vol. 33, no. 4, pp. 1034–1040, 2020.

[43] H. Xu, N. Usuyama, J. Bagga, S. Zhang, R. Rao, T. Naumann, C. Wong, Z. Gero, J. González, Y. Gu, Y. Xu, M. Wei, W. Wang, S. Ma, F. Wei, J. Yang, C. Li, J. Gao, J. Rosemon, T. Bower, S. Lee, R. Weerasinghe, B. J. Wright, A. Robicsek, B. Piening, C. Bifulco, S. Wang, and H. Poon, "A whole-slide foundation model for digital pathology from real-world data," *Nature*, vol. 630, pp. 181–188, June 2024.

[44] I. Nordrum, B. Engum, E. Rinde, A. Finseth, H. Ericsson, M. Kearney, H. Stalsberg, and T. J. Eide, "Remote frozen section service: A telepathology project in northern norway," *Human Pathology*, vol. 22, no. 6, pp. 514–518, 1991.

[45] C. Daniel, M. G. Rojo, J. Klossa, V. Della Mea, D. Booker, B. A. Beckwith, and T. Schrader, "Standardizing the use of whole slide images in digital pathology," *Computerized Medical Imaging and Graphics: The Official Journal of the Computerized Medical Imaging Society*, vol. 35, no. 7-8, pp. 496–505, 2011.

[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023.

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

[48] R. Jiang, X. Yin, P. Yang, L. Cheng, J. Hu, J. Yang, Y. Wang, X. Fu, L. Shang, L. Li, W. Lin, H. Zhou, F. Chen, X. Zhang, Z. Hu, and H. Lv, "A transformer-based weakly supervised computational pathology method for clinical-grade diagnosis and molecular marker discovery of gliomas," *Nature Machine Intelligence*, vol. 6, pp. 1–16, 07 2024.

[49] E. Vorontsov, A. Bozkurt, A. Casson, G. Shaikovski, M. Zelechowski, K. Severson, E. Zimmermann, J. Hall, N. Tenenholtz, N. Fusi, E. Yang, P. Mathieu, A. van Eck, D. Lee, J. Viret, E. Robert, Y. K. Wang, J. D. Kunz, M. C. H. Lee, J. H. Bernhard, R. A. Godrich, G. Oakley, E. Millar, M. Hanna, H. Wen, J. A. Retamero, W. A. Moye, R. Yousfi, C. Kanan, D. S. Klimstra, B. Rothrock, S. Liu, and T. J. Fuchs, "A foundation model for clinical-grade computational pathology and rare cancers detection," *Nature Medicine*, vol. 30, no. 10, pp. 2924–2935, 2024.

[50] R. J. Chen, T. Ding, M. Y. Lu, D. F. K. Williamson, G. Jaume, A. H. Song, B. Chen, A. Zhang, D. Shao, M. Shaban, M. Williams, L. Oldenburg, L. L. Weishaupt, J. J. Wang, A. Vaidya, L. P. Le, G. Gerber, S. Sahai, W. Williams, and F. Mahmood, "Towards a general-purpose foundation model for computational pathology," *Nature Medicine*, vol. 30, no. 3, pp. 850–862, 2024.

[51] H. Xu, N. Usuyama, J. Bagga, S. Zhang, R. Rao, T. Naumann, C. Wong, Z. Gero, J. González, Y. Gu, Y. Xu, M. Wei, W. Wang, S. Ma, F. Wei, J. Yang, C. Li, J. Gao, J. Rosemon, T. Bower, S. Lee, R. Weerasinghe, B. J. Wright, A. Robicsek, B. Piening, C. Bifulco, S. Wang, and H. Poon, "A whole-slide foundation model for digital pathology from real-world data," *Nature*, 2024.

[52] Bioptimus, "H-optimus-1," 2025.

[53] M. Kang, H. Song, S. Park, D. Yoo, and S. Pereira, "Benchmarking self-supervised learning on diverse pathology datasets," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3344–3354, June 2023.

[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[55] Y. Çelik and M. Karabatak, "Extracting low dimensional representations from large size whole slide images using deep convolutional autoencoders," *Expert Systems*, vol. 40, no. 4, p. e12819, 2023.