# A Closer Look at Weak Supervision's Limitations in WSI Recurrence Score Prediction

by

Namitha Guruprasad

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science
University of Alberta

# Abstract

Histological examination and derived ancillary testing remain the gold standard for breast cancer diagnosis, prognosis assessment and treatment guidance. Currently, a commercial molecular signature test OncotypeDX®, based on RNA quantitation and providing a recurrence score (RS) ranging from 0 to 100, is routinely utilized for luminal breast cancers (the largest sub-type group of breast cancers) to predict the probabilities of response to chemotherapy and disease recurrence. We attempt to predict RS using digital pathology and Weakly Supervised (WS) attention-based models. In tissue samples, the malignant component is haphazardly admixed with the non-malignant component in variable proportions. This represents a challenge for WS attention-based models to identify high-valued diagnostic/prognostic areas within whole slide images (WSIs). To address this, we propose an interactive, supervised approach with a human in the middle by creating a user-friendly Graphical User Interface (GUI) that allows an expert pathologist to annotate heatmaps generated by any WS attention-based model. We aim to enhance the model's learning capabilities and performance by incorporating the feedback from the GUI as *expected scores* in the successive training process. We train WS attention-based models like **CLAM** (Clustering-constrained Attention Multiple Instance Learning) [1] and **TransMIL** (Transformer based Correlated Multiple Instance Learning) [2] on our in-house dataset before and after the expert feedback. We observe an improvement in RS prediction after retraining both models with the pathologist's annotation- a 5% rise in validation-test AUC and 4% in validation-test accuracy for CLAM and a 4.5% increase in validation-test AUC and 3% in validation-test accuracy for TransMIL. We

analyze the generated heatmaps and observe how additional supervision from a domain expert enhances the learning capacity of the models. We notice an improvement in cosine similarity between the pathologist's GUI-based attention scores and trained models' attention maps after feedback - 5% and 10% increase for CLAM and TransMIL, respectively. The implementation of the proposed approach and the dataset is available for download[1]. Our adaptive, interactive system harmonizes attention scores with expert intuition and instills higher confidence in the system's predictions. This study establishes a potent synergy between AI and expert collaboration, addressing the constraints of WS by enhancing the discrimination of diagnostic features and making an effort to generate predictions according to clinical diagnostic norms.

[1] https://github.com/nam1410/RS_prediction.git

# Preface

This thesis draws its foundation from the contribution for presentation at The IEEE International Conference on Bioinformatics & Biomedicine (IEEE BIBM 2023), where it awaits review by peers in the field.

*To everyone who has steadied my course throughout this journey, your presence has been my anchor and compass. Your unwavering influence has shaped not only my path but also the tapestry of my understanding.*

*Misery grows, I'm close to giving up hope .....*

*.....Holding on to my life, afraid to let it go;*

*Where toxic thoughts in my head would unify .....*

*..... Surviving the madness, the world's an asylum;*

*Where I could crumble under pressure or turn to a diamond .....*

*..... Every day I was sunk in despair;*

*Only wheezing as I breathe through the pungent air;*

*Carried the weight on my shoulders, hence the hunch that I bear;*

*..... I guess pain is a fraction of what trouble evokes;*

*Success tastes sweetest for those who struggle the most .....*

*- BrodhaV (All Divine)*

# Acknowledgements

I am grateful for the invaluable guidance and unwavering support of Dr. Nilanjan Ray, Dr. Gilbert Bigras and Amir Akbarnejad. Their profound insights and mentorship have been pivotal in shaping the trajectory of this thesis.

Furthermore, I would like to thank DynaLIFE Medical Laboratory for their assistance in the slide scanning process.

It is also imperative to acknowledge that this undertaking has been made possible through the support provided by Mitacs Accelerate and Cross Cancer Institute.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 The WHERE problem in Whole Slide Images (WSI)

Breast cancer (BC) is a pressing global concern, given its estimated fatality count of 43,700 projected for 2023 in the USA. Hematoxylin and eosin (H&E) histological examination of needle core biopsy remains the routinely performed gold standard method for BC diagnosis. Pathologists have excelled in precise cancer diagnosis for many years, but prognosis prediction presents a different challenge. Breast cancer comprises four subtypes (as shown in **Fig. 1.1**): Luminal A (including normal-like), associated with a good prognosis, and Luminal B, HER-2 enriched and Triple negative, associated with bad prognoses, necessitating aggressive chemotherapy. Pathologists can readily identify HER-2 and Triple Negative subtypes using in-house ancillary tests like H&E and Immunohistochemistry (IHC). However, distinguishing Luminal A from Luminal B, representing 65% of all BC cases, remains elusive. This thesis focuses on the distinction between Luminal A and Luminal B. Pathologists have attempted to classify them into grades where Grade 1 corresponds to Luminal A (good prognosis), Grade 3 to Luminal B (a bad prognosis), and Grade 2 remains uncertain (as shown in **Fig. 1.1**). However, this grading system suffers from poor inter-observer reproducibility, making it unreliable for treatment decisions. Nowadays, a costly commercial test called OncotypeDX® provides a Recurrence Score

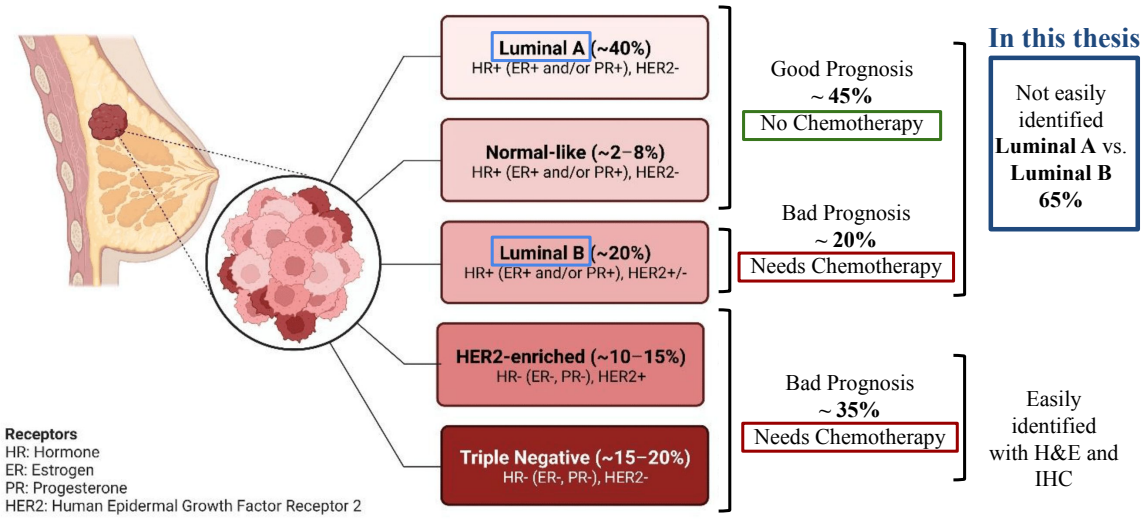



Figure 1.1: Subtypes of Breast Cancer (BC)

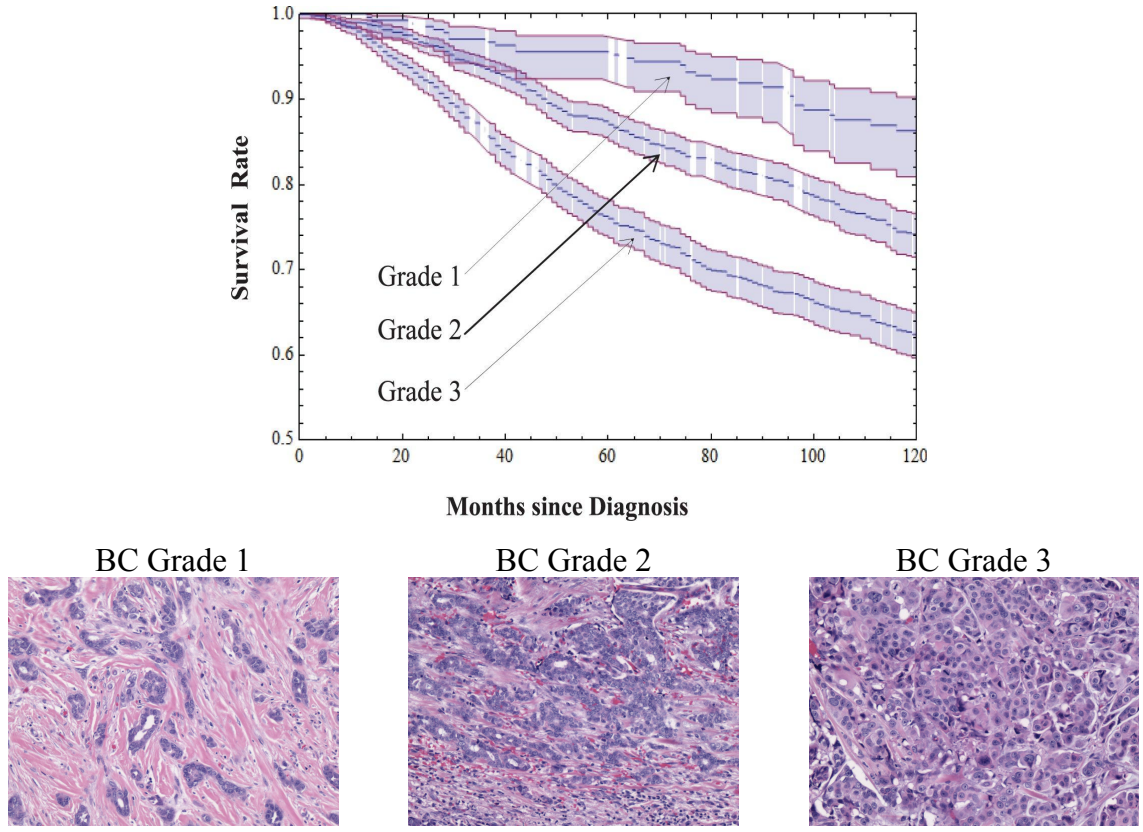



Figure 1.2: Grades of Breast Cancer (BC)

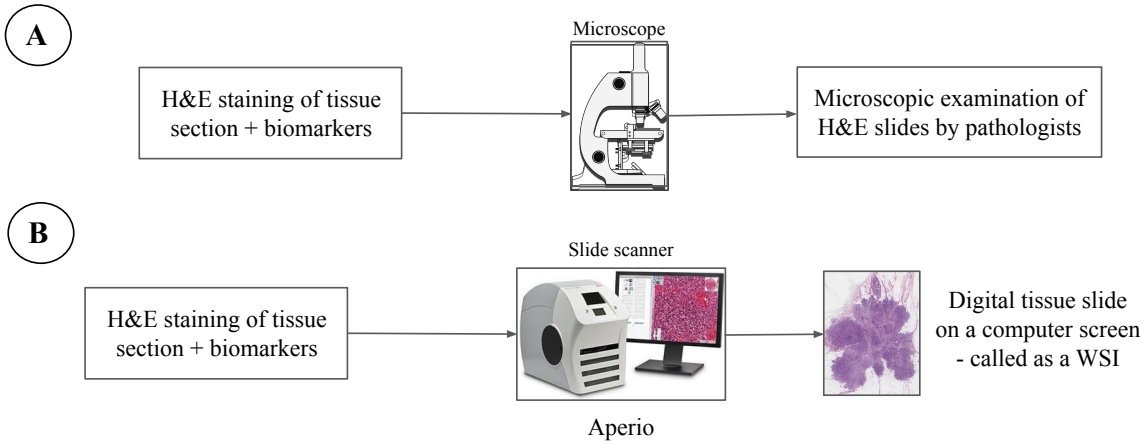



Figure 1.3: Digitization of tissue slides - Whole Slide Imaging (WSI)

(RS) to aid in prognosis prediction. **We aim to predict the OncotypeDX® score by leveraging the prognostic information within BC morphology**. RS, ranging from **0 to 100**, is based on expression levels of 16 genes normalized to 5 reference genes. An RS value of **25 and lower** defines the risk of recurrence to be ***low***, whereas an RS value of **higher than 25** defines a ***high*** risk and decides to add chemotherapy to the treatment plan [3].

Traditionally, pathologists manually examine tissue sections stained with H&E and other biomarkers through microscopes. The engineering of reliable and fast scanners has facilitated the digitization of entire tissue slides, resulting in large images known as Whole Slide Images (WSIs). However, analyzing WSIs presents a formidable challenge (as shown in **Fig. 1.3**). These images are gigapixels in size, about 50,000 times larger than images in datasets like ImageNet with an average size of 469-by-387 pixels. While identifying objects like a bald eagle by Deep Learning models (as shown in **Fig. 1.4**) is often less complicated, it becomes difficult to pinpoint tumorous and non-tumorous components in WSIs. It indicates that the crucial diagnostic information is scattered and sparsely distributed, comprising a variable proportion of the total image area. It is called the *WHERE problem*, a situation arising from disproportional distribution and varying sizes of tumour regions.

The crux of the *WHERE problem* lies where these diagnostically significant tumour

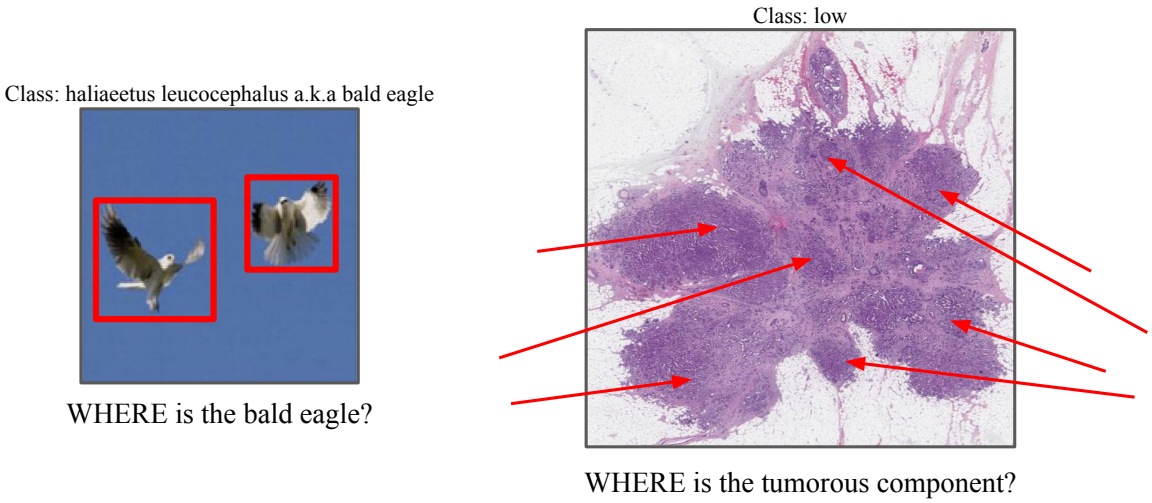



Figure 1.4: The **WHERE** problem

areas often occupy an undetermined fraction of the WSI, along with an undetermined fraction of non-tumoural areas (surrounding normal epithelial structures, normal stromal components - fat, vessels, muscle, etc) and also inflammatory areas triggered or not by the tumoural process. This dynamic poses a distinct challenge: discerning the relevant malignant content from the surrounding non-malignant content is in some cases akin to searching for a **needle in a haystack**. Addressing the *WHERE* challenge requires an understanding of the spatial distribution of tumour features in the WSI. Along with this, the developing algorithms should be capable of discriminating between subtle variations in tissue patterns across the entire WSI.

Obtaining precise delineations from pathologists takes time, thus leading to Weak Supervision (WS), where weak labels are assigned manually or based on empirical results. Multiple Instance Learning (MIL) partitions a WSI into patches, allowing confident labelling of the entire bag of patches (instances) when individual instance labelling in a bag is uncertain or expensive (as shown in **Fig. 1.5**). An aggregation mechanism synthesizes information from these patches. To identify the **patches with high diagnostic significance**, an attention mechanism becomes crucial for highlighting the relevance of objects relevant to predictive tasks. In natural language processing and image analysis, the attention mechanism has demonstrated its effec-

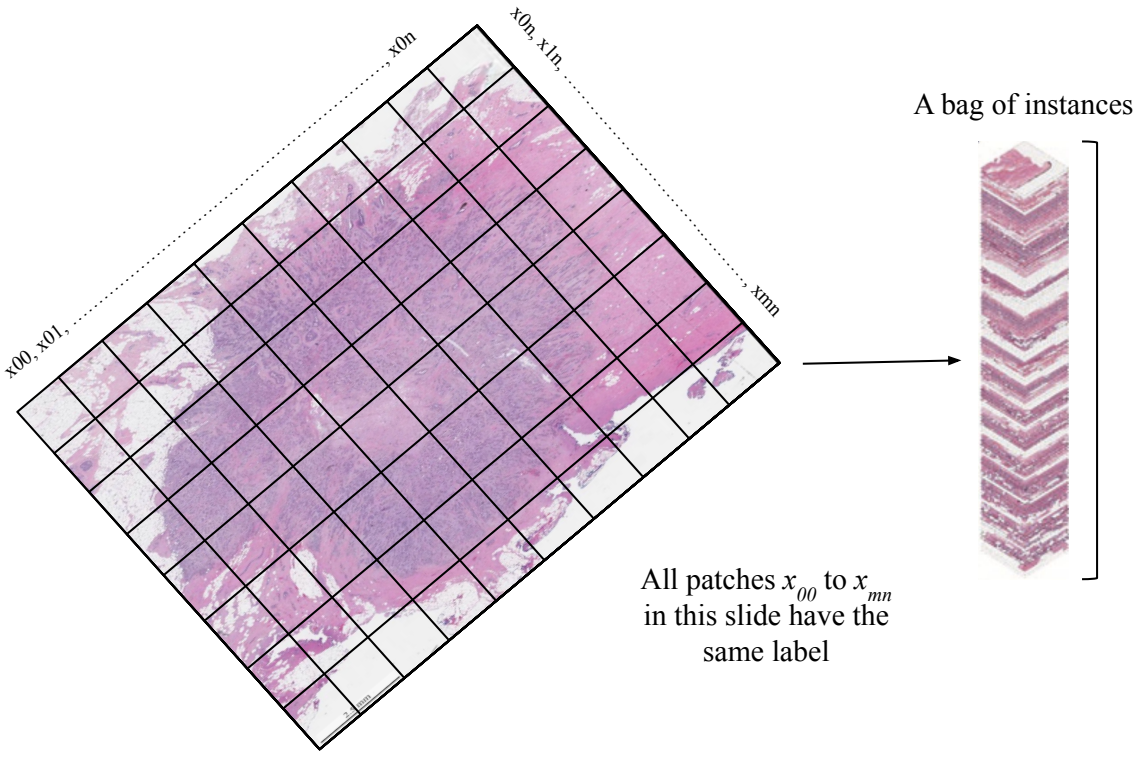
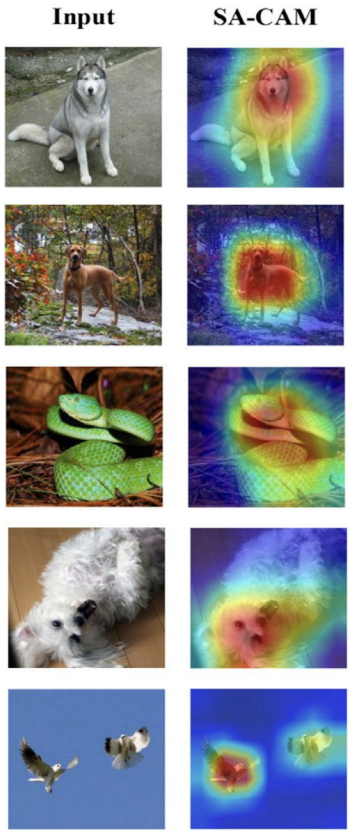


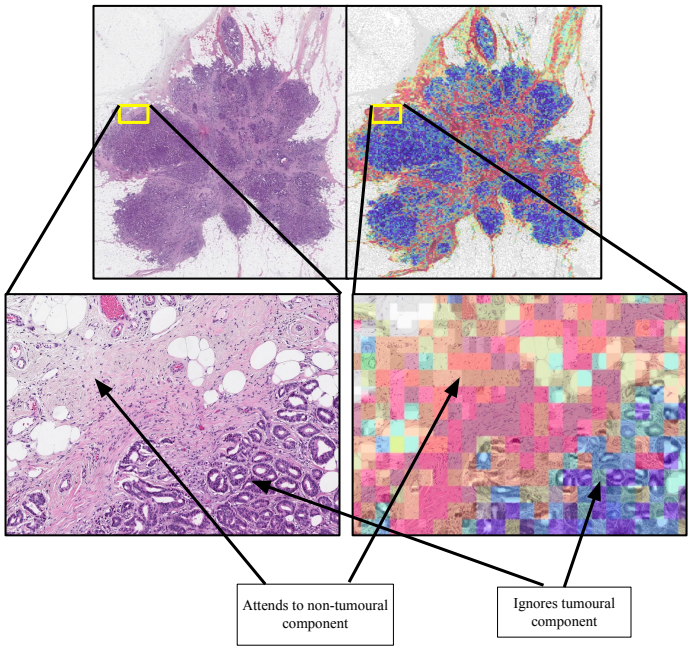

Figure 1.5: Multiple Instance Learning (MIL)

tiveness in singling out critical elements that impact decision-making. For instance, the Self-Attention Class Activation Map (SACAM) [4] (as shown in **Fig. 1.6**) identifies objects like animals within images. The success of such mechanisms in datasets like ImageNet has enticed researchers to explore their applicability in diverse domains, including medical imaging. Consequently, the question emerges: **Can self-attention enhance the prediction of cancer recurrence likelihood?**

Weakly supervised models like **CLAM** (Clustering-constrained Attention Multiple Instance Learning) [1] and **TransMIL** (Transformer based Correlated Multiple Instance Learning) [2] exhibit strong performance (AUC $>$ 0.95) in primary predictions involving metastatic tissues, especially in benchmark datasets like TCGA (The Cancer Genome Atlas Program). Notably, the WSI size within this dataset typically averages around 40k-by-40k pixels, in contrast to recurrence score datasets with slides around 100k-by-100k pixels. Metastasis evaluation primarily focuses on assessing the spread of cancer to different regions within the body, while recurrence score predic-




(a) SACAM




(b) The visual depiction illustrates the disparity in attention heatmap coherence within a sub-typing context (**left:** H&E image; **right:** corresponding attention heatmap); The heatmap portrayal of a WSI depicts an attention pattern discordant with the diagnostic expectations, where the prominent red nodule denoting the non-malignant component is attended to by the model (*negative picture* of the attention)

Figure 1.6: Attention based mechanism

tion pertains to the likelihood of cancer re-emergence. Accurately predicting prognosis and treatment response remains an ongoing challenge, particularly for subtyping tasks that involve distinguishing between high and low-risk cases. Predicting recurrence introduces a higher degree of complexity and is less straightforward, rendering it a more formidable task. For instance, *when we apply CLAM and TransMIL to our recurrence score dataset, they frequently and unpredictably favour non-malignant components while overlooking essential discriminative features linked to the malignant regions.* The generated heatmap deviates from the expected clinical diagnostic results, leading to erroneous false-positive and false-negative outcomes (as shown in **Fig. 1.6**) called the **negative picture of attention**. This seemingly counterintu-

itive behaviour of WS attention models like CLAM and TransMIL underscores the critical **need for domain experts to provide guidance**. This guidance often involves offering feedback to indicate where the model's predictions align with the ground truth and where it deviates.

In medical imaging, **human supervision** has been in practice, necessitating the integration of manually labelled data into the model training process. For instance, [5] and citele2019pancreatic, collect annotations through crowdsourcing or manual markings. However, these fixed annotations would not align optimally with the model's evolving learning goals, potentially impeding its capacity to detect intricate patterns.

## 1.2 Objectives

In light of the **negative picture of attention** and the essential role of **human guidance**, we endeavour to address one of the most challenging tasks: **predicting recurrence score**. In this context, we leverage the expertise of domain specialists to solely guide the WS model in achieving a more refined differentiation between malignant and non-malignant components. Our research inquiries are as follows:

- Can we devise an interactive feedback collection system for annotating a WS attention-based model's heatmaps? Can we minimize the time and effort required for such annotations?

  To address this, we develop a Graphical User Interface (GUI) by integrating Openseadragon[1] with Openslide[2]. This enables a pathologist to delineate free-bounding boxes of any size on any attention-model's heatmap of a WSI.

- How can we effectively incorporate the pathologist's feedback from the GUI into the attention model's re-training process and simultaneously enhance its learning of relevant features and performance?

[1]https://openseadragon.github.io/
[2]https://openslide.org/

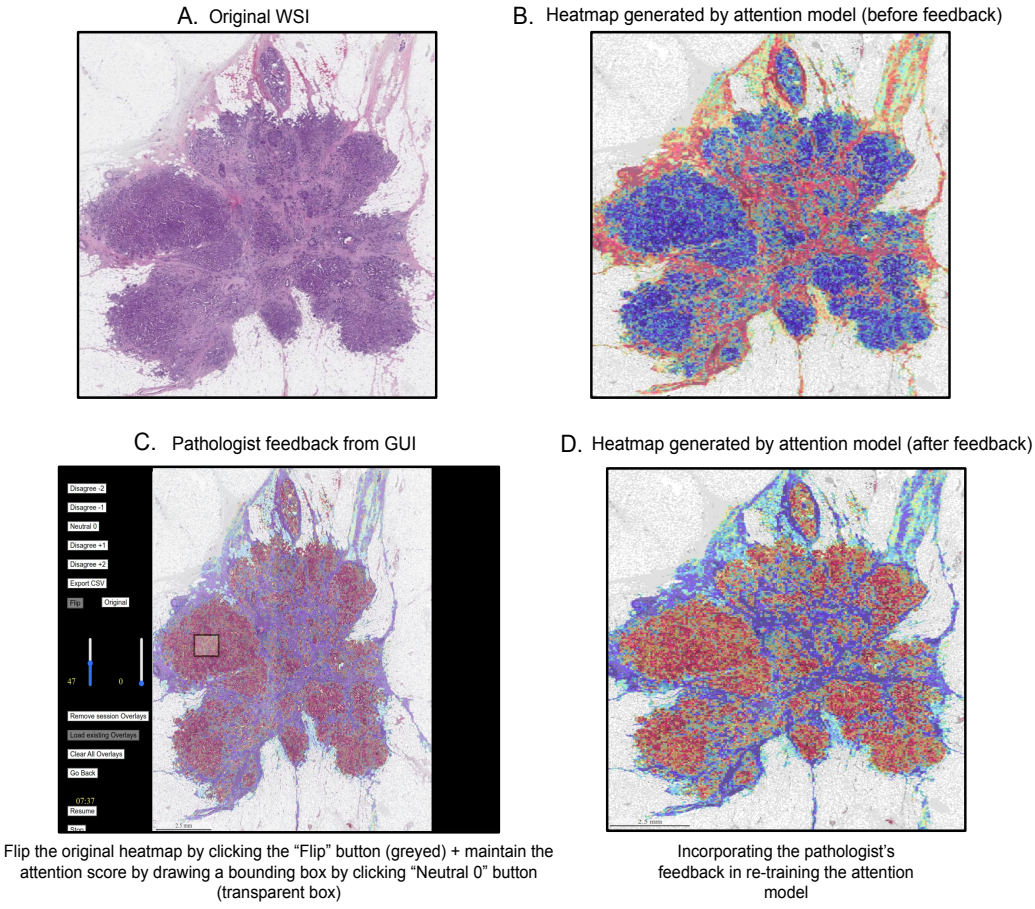



Figure 1.7: Pathologist provides feedback using the GUI.

To address this, we capture the patch attention scores from the initial heatmaps (before annotation), then perform a spline interpolation with the pathologist's feedback information. In other words, this approach modifies the attention scores at every region on a WSI based on the reason (to increase, decrease or maintain attention scores) for drawing the bounding box at that region, calling them the *expected scores*. We run the attention model by incorporating a Mean Squared Error (MSE) loss to penalize the difference between *expected scores* and the new attention scores generated by the model in the ongoing training cycle (*current attention scores*). We also observe the resulting changes in the heatmaps before and after the feedback from the expert and calculate their cosine similarity for every WSI.

To summarize, we align the model's attention scores to generate heatmaps more consistent with what an expert would anticipate, thus instilling greater confidence in its predictions.

Our study is grounded in vital supervision through expert annotations. By doing so, we rectify the limitations of WS attention-based models by fostering their collaboration with the pathologists to improve model prediction according to the clinical diagnostic standards.

## 1.3 Outline

The subsequent sections of this thesis are structured as follows. The initial portion of **Chapter 2** delves into the foundational aspects of Whole Slide Imaging, encompassing the associated challenges and prospects. The subsequent segment of **Chapter 2** sheds light on the pertinent literature and highlights the existing gaps in the field.

Moving forward, **Chapter 3** of the thesis explains the extraction of attention scores from convolution and transformer-based architectures to generate heatmaps that serve as the basis for annotation.

**Chapter 4** explores our GUI, which enables pathologists to annotate the generated heatmaps and effectively process their annotations. Additionally, this chapter elucidates the methodologies applied to incorporate such annotations in the model's re-training process.

**Chapter 5** offers an in-depth account of the experimental setup, detailing the specifics of our approach and presenting the obtained results.

Finally, **Chapter 6** serves as the concluding segment of the thesis, summarizing the key insights and contributions derived from our work and encapsulating unexplored directions for future research.

# Chapter 2

# Background

We elaborate on the concepts essential for understanding the Whole Slide Image (WSI) analysis. We segue into an in-depth exploration of the pertinent literature to shed light on WSI diagnosis.

## 2.1 A note on recurrence score

Predicting cancer recurrence and response to therapy holds paramount significance in the diagnostic follow-up process. While achieving accurate cancer diagnosis is nowadays properly managed by expert pathologists, predicting accurate prognosis and response to therapy remains an ongoing challenge within the healthcare domain. Since 2005, as we had seen in **Section 1.1** the routinely performed biological biomarkers (ER, PR, HER2/neu) have sub-typed Breast Cancer (BC) as Luminal, Triple Negative and HER2/neu, permitting significant therapeutic standardization. However further sub-classification of the Luminal BC sub-group has been, until recently, a challenge. Indeed, Luminal BC patients form the largest subset of BC for whom the decision of adding chemotherapy in their treatment has been historically a difficult clinical decision. There are two Luminal BC subsets: Luminal A which does require chemotherapy and Luminal B which does not. Since 5 years, this prediction is facilitated by the utilization of the **Recurrence Score (RS)** that guides oncologists. This test employs *RT-qPCR* to gauge the activity of a cluster of 21 genes (16 breast

cancer-related normalized to 5 reference housekeeping genes). The 16 genes are represented as follows: a proliferation group (Ki-67, STK15, Survivin, Cyclin B1 and MYBL2), an Estrogen group (ER, PGR, Bcl2 and SCUBE2), a HER2 group (GRB7 and HER2), an invasion group (Stromelysin 3 and Cathepsin L2) and the individual genes CD68, BAG1 and GSTM1. The RS is computed using the following linear equation:

$$RS = 0.47 \times \text{HER2 group score} - 0.34 \times \text{Estrogen group score} +$$
$$1.04 \times \text{Proliferation group score} + 0.10 \times \text{Invasion group score} +$$
$$0.05 \times \text{CD68} - 0.08 \times \text{GSTM1} - 0.07 \times \text{BAG1}$$

Nowadays all BC patients who have high HER2 expression (typically BC with demonstrated HER2 gene amplification) are not tested with OncotypeDX®. Therefore the RS is mostly determined by the proliferation and the Estrogen groups.

Of interest, RS is based on biological data which can be derived from 4 in-house immuno-histochemistry (IHC) assays (ER, PR, HER2/neu and KI-67) routinely performed in Canadian laboratories. The success of this OncotypeDX® RS is explained by the poor inter-laboratory reproducibility of these 4 IHCs. RS spans from 0 to 100 - an RS value of 25 or lower denotes a low risk of recurrence; an RS value of higher than 25 signifies an elevated risk, potentially warranting chemotherapy incorporation into the treatment strategy. Several retrospective and prospective studies validate this test and its clinical utility. [6] demonstrates a correlation between RS and disease-free survival among patients with estrogen receptor (ER)-positive, HER2-negative, node-negative breast cancer, specifically within the NSABP B-14 trial. Notably, TailorX trial [3] recently confirms that for HR-positive, HER2-negative early-stage breast cancer patients aged over 50 years and those aged below 50 with an RS $\leq$ 15, chemotherapy may be omitted. This testament of the OncotypeDX® test is grounded in Level 1A evidence and garnered endorsement within prominent international clinical guideline recommendations, including those set forth by the American

Society of Clinical Oncology (ASCO). While OncotypeDX® RS is considered the gold standard, this test costs 100 times the cost of the 4 IHCs which economically impacts Canadian laboratories. It is suggested that AI can palliate the problem of reproducibility introduced by human readouts [7].

## 2.2 First principles: Whole Slide Imaging

In conjunction with the above genomic tests, pathologists engage in microscopic evaluations of tissue sections stained with hematoxylin and eosin (H&E), as well as utilize diverse biomarkers in specialized IHC and Immunofluorescence (IF) examinations as part of their diagnostic protocols. These practices facilitate the visual identification of regions, a process underpinned by the pathologist's extensive medical expertise. The evolution of computer processing power, data transfer speeds and storage solutions has transferred the act of scrutinizing glass microscope slides to computer monitors through dedicated scanners with software tools. This transformative process of converting histology slides into a digital format yields huge digital images termed Whole Slide Images (WSIs). These WSIs, often attaining gigapixel dimensions (around 100K-by-100K pixels), adhere rigorously to stringent diagnostic quality benchmarks. Upon accessing the digital file, pathologists navigate, zoom, and observe spatially on a computer screen, mimicking the conventional light microscope's functionality. This approach allows them to inspect regions of interest, identifying pertinent features or patterns like traditional microscopy techniques.

### 2.2.1 Scanners

Digital slide scanners exhibit variability in their features and capabilities, encompassing factors like scanning capacity (ranging between 100 and 200 slides in a single batch), availability of objectives (20x or 40x magnification), and image resolution (typically falling within the range of 0.25 to 0.5$\mu$m per pixel) [8]. Most histopathological scanned images are acquired using bright field light microscopy with the 40x

objective, which is widely accepted as the standard level of magnification for digital slides and well-suited for a wide array of analyses. For example, consider a WSI scanned at 40x magnification, featuring an image resolution of approximately 0.25$\mu$m per pixel and a 24-bit colour depth. In such a case, the volume of data representing a $1mm^2$ slide area amounts to a staggering 384 million bits, resulting in a file size of around 48 MB without implementing additional strategies to optimize data management. This data size increases further when accounting for the entire slide area or the inclusion of multiple z-planes (only considered for cytological samples), thereby potentially placing demands on bandwidth and storage resources [9, 10].

### 2.2.2 Tiled pyramid

The utilization of memory-friendly compression techniques like JPEG, JPEG 2000, or LZW leads to substantial loss of digital data during the conversion process, often resulting in irretrievable information loss. Given that a WSI frequently surpasses 1 GB in size, loading such files into memory for display and navigation becomes impractical. In such scenarios, a reciprocal connection emerges between the image's scale and field of view. For a larger field of view, resolution becomes constrained by the monitor, necessitating the loading at the highest resolution. Conversely, for a smaller field of view, the entire image need not be loaded when examining the tissue at a relatively high magnification. Hence, WSIs are stored at multiple resolutions, establishing a streamlined approach for image loading and enabling more efficient image delivery. For instance, an Aperio Scanscope scanner might acquire a sample WSI at 40x magnification, accompanied by downsampled versions at 10x, 2.5x, and a thumbnail representation encapsulating the entire tissue within a 1-megapixel frame. This stratified, multi-resolution depiction is known as the **tiled pyramid** (as depicted in **Fig. 2.1**), offering improved data throughput by precomputing lower-resolution renditions of WSIs. Most WSI files are of the Tagged Image File (TIF) or Aperio ScanScope Virtual Slide (SVS) format [11].

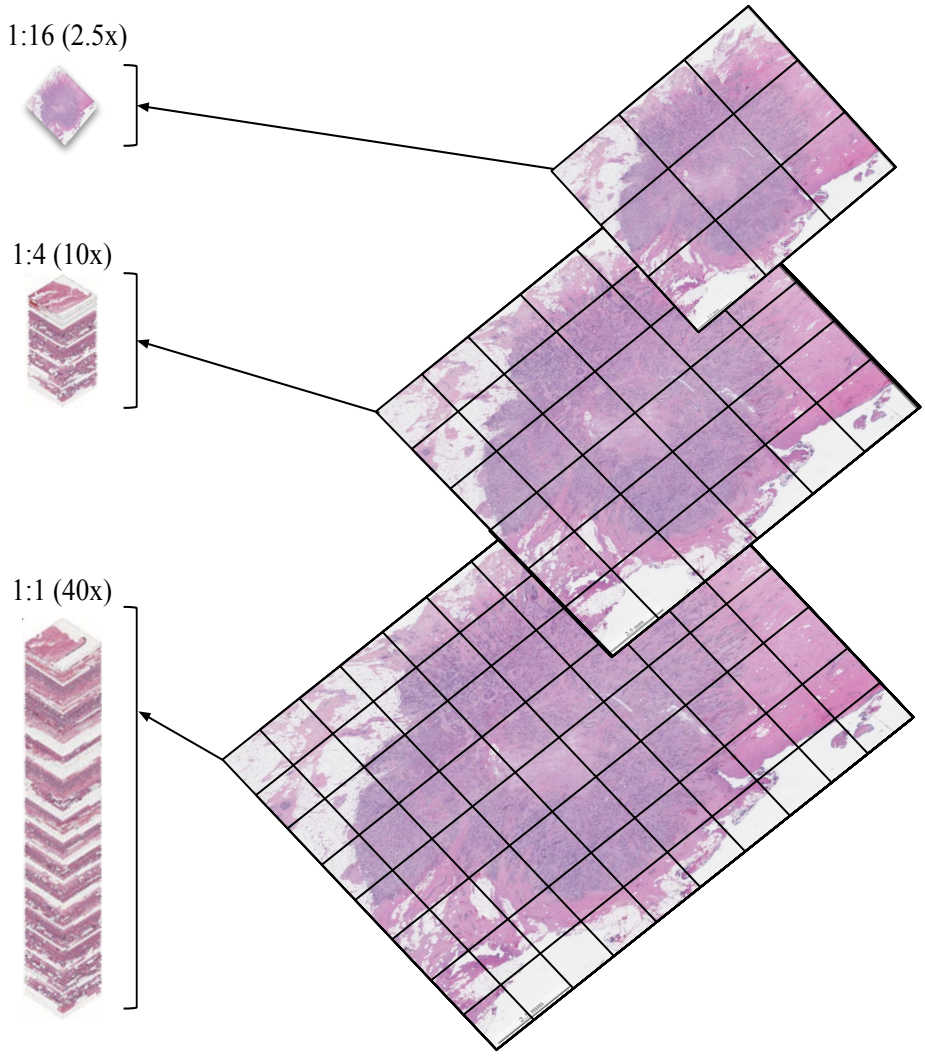



Figure 2.1: **WSI pyramid** - Each level in the pyramid represents a different pre-calculated zoom level. The base level represents the full-resolution image. Moving up the pyramid, each subsequent layer is digitally sub-sampled to create a lower magnification view, ending at the lowest magnification image at the apex of the pyramid. Each level is further divided into small tiles of a fixed size (not shown to scale). Since the tiles are of a fixed size, the base level is composed of the largest number of tiles, with each subsequent level being composed of fewer and fewer tiles.

### 2.2.3 Annotation and labels

Advanced slide scanners have integrated capabilities for tissue identification and autofocusing. Moreover, WSI is exceptionally well-suited for applying algorithms to aid in the analysis of digital tissue images for pathologists. Diagnostic tools featuring algorithms extract pertinent parameters from WSI scans by comparing tissue sections or even individual pixel colours against predefined diagnostic standards [12]. For such tools, pathologists play a critical role in evaluating structural, textural, and morphological markers to identify tumour regions for a specific clinical task or issue. Depending on the task's requirements, implementations may be variable, encompassing point annotations (identifying the centroid of the pathology marker), shape annotations (defining a predefined shape around the pathology marker), or intricate outline annotations (precision segmenting of the pathology marker).

In contrast to the above annotation which assigns labels to distinct morphological attributes, categorization enables assigning an entire WSI to a diagnostic category (for instance, *high*-risk or *low*-risk in terms of RS). Noteworthy examples of some diagnostic labels encompass disease subtype, grade, administered medications, and survival rates. These labels are derived from patient records or established by domain experts who review the clinical data.

## 2.3 Opportunities

Fundamental aspects like diagnosis and histological grading are pivotal in disease assessment and treatment strategy. It entails identifying tissue patterns and cellular attributes linked to diverse pathologies. Traditionally, pathologists rely on recognizing tissue patterns based on cellular morphology, cell distribution, and architectural configurations to achieve diagnoses. Once a cancer diagnostic is established, the pathologist attempts to predict the associated prognosis by gauging the extent of abnormality, the degree of cellular differentiation, and other histological traits to assign

a histological grade to the disease, based on an ordinal scale of measurements. This assessment is performed on the malignant tissue which is found in specific areas of a the WSI. A human expert can find the region of interest quickly but it remains an algorithmic challenge given the huge size and potential random informative content of a WSI.

The subsequent section delves into the literature on the necessity for advanced algorithms to facilitate the identification of diagnostic regions within WSIs. Starting with a discussion on traditional machine learning algorithms, we explore weak and strong supervision strategies and elucidate the rationale behind the demand for interactive feedback from pathologists.

## 2.4 Literature Review

### Traditional Diagnostic Tools

Feature extraction holds prominence, as the efficacy of algorithms directly corresponds to the quality of features derived from input images. Traditional techniques for feature extraction, such as Local Binary Pattern (LBP) [13] and Local Phase Quantization [14], are assessed in conjunction with diverse classifiers, including 1 Nearest Neighbor (1-NN) [15], Quadratic Linear Analysis (QDA) [16], Support Vector Machine (SVM) [17], and Random Forest (RF) [18]. The investigation indicates that SVM exhibited commendable performance on low-resolution images when employing fractal dimension as a feature descriptor [19]. Further, following the extraction of features from the histopathological images, the approach in [20] classifies them into *IDC-* and *IDC+* categories. However, the results exhibit some inconsistencies due to the wholesome reliance of these techniques on the quality of features obtained through various feature descriptors [21].

Interactive diagnostic tools like Cellprofiler [22, 23] require partial manual delineations for extracting features at the pixel level (Kirsch, Haralick features, HoG, LBP), object level (nuclei spatial dependency - Voronoi, Minimum Spanning Tree, De-

launay triangulation), and semantic level (graph embeddings) which go into classical classifiers such as Logistic Regression, Support Vector Machine (SVM), and Decision Trees as inputs. Although these software-based analytical approaches acquire quantitative, reproducible, and objective data, the nature of tissue sections affects the efficacy of such software. It is due to their incapability to identify irrelevant regions that may be overlooked by the human eye, such as edge effects in staining, tissue folding, and variations in section thickness, all of which can lead to erroneous results [24]. The limitations of these traditional approaches prompt the exploration of alternative methodologies that can better accommodate the complexity and nuances inherent in Whole Slide Images (WSIs) within the field of pathology.

### 2.4.1 Weak Supervision (WS)

The adoption of Weak Supervision in WSI analysis harnesses the potential of Deep Learning to navigate the challenges presented by the intricate nature of WSIs. The emergence of Deep Learning algorithms, renowned for their exceptional performance on platforms like ImageNet [25], opens a novel avenue for WSI analysis. In tackling the immense scale of WSIs, the Multiple Instance Learning (MIL) strategy entails partitioning the multi-gigapixel slide (bag) into smaller patches (instances) that the machine can handle and subsequently employs an aggregation mechanism to consolidate all the information from these patches into a final prediction. Since it leverages only slide-level labels during the learning process, this approach is called Weakly Supervised (WS) learning.

WS models enable comprehensive exploration of the entire tissue slide with its constituent patches. The method proposed by [26] assigns weights or scores to instances that wield influence over the final prediction through an aggregation mechanism to find the key instances of high diagnostic value. However, this attention-based technique exhibits overfitting issues due to a restricted number of slides and a bias towards benign samples [27] in a binary classification between malignant and benign tissue

slides. It also lacks explicit patch relation modelling because it uses a multi-layer perceptron to predict attention scores for each patch. Another approach, [28], implements RNN-based aggregation for patient-level prediction, involving patch-level training and top-k instance selection. The inherent limitation of representing patch features as a 1-dimensional sequence hinders its ability to capture the 2-dimensional spatial positions within WSIs.

The above WS methods do not directly apply to multiclass tissue subtyping problems (high-grade and low-grade cancer types, i.e., *high* vs. *low* RS), especially in cases where benign tissue slides are unavailable. As a response, the attention-based textbf-CLAM (Clustering-constrained Attention Multiple Instance Learning) [1]framework concentrates on identifying regions relevant for predictive determinations by instance-level clustering to generate interpretable heatmaps for multiclass classification tasks. The widely adopted assumption of instances being independent and identically distributed (i.i.d.) does not suitably apply in pathology, where contextual information and the relationship between different tissue regions are essential for making diagnostic judgments. In a parallel advancement, [29] introduces the Vision Transformer (ViT), which directly applies to different image patches arranged sequentially for classification tasks. As a response, and **TransMIL** (Transformer based Correlated Multiple Instance Learning) [2] leverages the self-attention mechanism to encode the mutual correlations between instances with interpatch dependencies and quantifies the attention scores of each instance contributing to the bag classification.

### 2.4.2 Strong Supervision by Pathologist's input

The need for an extensive training dataset becomes noteworthy when WS methods do not grasp the features in an effective way. The intricate variations within distinct tissue types, the diversity of histopathological patterns, and the subtle differentiations between malignant and non-malignant regions emphasize the necessity of a human-guided training-error learning strategy.

Such a strategy has success in other domains, such as radiology for Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) images. Related to pathology, for example, in [30], human-provided scribbles are transformed into pseudo-label maps (associated with confidence values) by a probability-modulated geodesic transform. Confidence values are allocated to pixels demonstrating large geodesic distances from the scribbles or displaying low network output probabilities, thus making it an iterative learning process. Approaches like [31, 32] employ eye tracking to observe the complexity of pathologists' diagnostic decision-making while observing WSIs. The attention patterns exhibited by pathologists are captured using a web-based digital microscope and subjected to distinctive diagnostic search patterns with scanning and focused examination. Given the large size and high resolution of WSIs, annotators sometimes resort to generating coarse annotations, as seen in [33]. This technique involves annotators outlining the approximate boundaries of cancerous regions at lower to intermediate magnifications. In contrast, the approach introduced by [5] requires annotators to place a limited number of points on each WSI at lower magnifications, thereby being cost-effective compared to intrinsic annotation methods. It incorporates a robust architecture distinguishing between various classes and subclasses of tissues under diagnosis. Similarly, bounding boxes are user-friendly supervision, requiring only a fraction of the annotation time compared to full pixel-wise annotations [34].

## 2.5 Gap in existing literature

WS models (in **Section 2.4.1**) are trained exclusively using slide-level labels but they often struggle to identify vital diagnostic features while inadvertently emphasizing irrelevant ones (*negative picture* of attention). Similar observations in seminal studies such as [5] and [35] suggest that the compromise in precision and reliability questions their clinical applicability.

Conversely, the Strong Supervision approach (in **Section 2.4.2**) involves upfront

annotations acquired through crowdsourcing. This static approach fails because traditional annotation methodologies confine the model within initial annotations, restricting adaptation to emerging insights. Bridging this gap necessitates accommodating the dynamic nature of WSIs and diagnostic models.

The subsequent chapters emphasize the integration of interactive expert feedback (Human supervision) to reduce the attention to non-malignant components (*negative picture of attention*) by the WS attention-based models.

# Chapter 3

# Attention scores

## 3.1 WS drawbacks - Counter-intuitive behaviour associated with tumoural regions

As mentioned in **Chapters 1** and **2**, in recent years, weakly supervised (WS) learning has garnered significant attention from researchers and has found application in pathology. We have seen before that the malignant component may only occupy a small portion of the slide, with most of the WSI potentially comprising non-malignant constituents. Moreover, the spatial arrangement of the malignancy within the WSI exhibit substantial variation from case to case. Attention-based pipelines like WS attention-based models like **CLAM** (Clustering-constrained Attention Multiple Instance Learning) [1] and **TransMIL** (Transformer based Correlated Multiple Instance Learning) [2] exhibit a counter-intuitive behaviour by emphasizing non-malignant components while neglecting the essential malignant ones and eventually misinterpretation. It means that some patterns within non-tumoural areas receive attention but overlook patterns within tumoural regions for the MIL-based method in subtyping due to the complex nature of WSIs.

The visual representation in **Fig. 3.1** illustrates issues with attention weights encountered in our initial attempt to classify *high* vs. *low* RS. In **Fig. 3.1A**, the heatmap for a WSI incorrectly classified as belonging to class *high* (actually *low*) displays a **negative picture** of attention (paying attention to non-tumoural com-

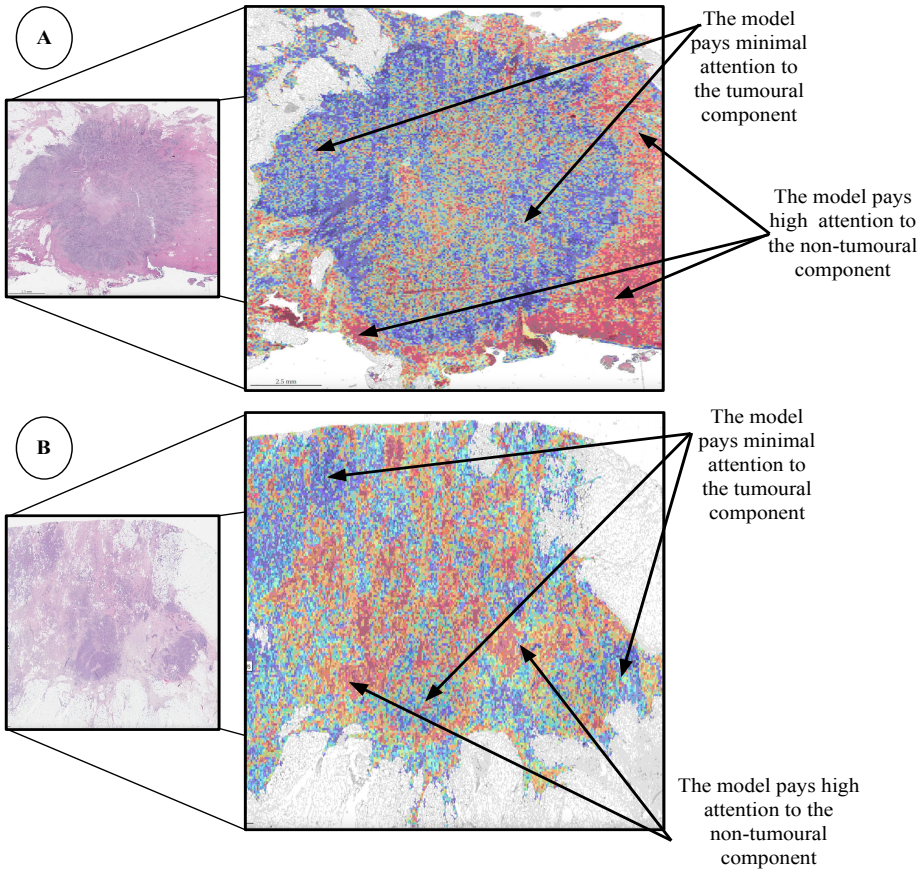



Figure 3.1: Visual representation of heatmap inconsistencies in a sub-typing problem (**left:** H&E image; **right:** corresponding attention heatmap)
**A**- Heatmap of a misclassified WSI (*high* class incorrectly labeled as *low*) displaying a *negative picture* of attention where the *blue nodule* representing the tumour is almost totally ignored by the model; **B**- Heatmap of a correctly classified WSI displaying another *negative picture* of attention. The heatmap appears fuzzy as the tumour itself is poorly defined in the corresponding H&E image
**NOTE:** The arrows represent the discrepancies in the heatmaps with respect to the attention paid by the model (targeting tumoural vs. non-tumoural components).

ponent), assuming that it is arising due to its incorrect classification. Conversely, **Fig. 3.1B** demonstrates the scenario where, although the WSI is correctly classified, the heatmap generated is another **negative picture** of attention. Irrespective of the classification results, WS attention models signal the persistence of non-malignant features originating from non-tumoural areas or artifacts. These artifacts encompass, for instance, air bubbles, tissue folding, and ink markers, which interfere with the proper representation of tumoural regions. Some might argue that this counterintuitive behaviour may be due to a limited number of WSIs. While expanding pathology datasets remains a potential solution, it is crucial to acknowledge the considerable time and memory required for curating thousands of WSIs, each corresponding to an individual patient. Currently, we focus solely on the available limited WSIs scenario. The initial heatmaps generated through WS attention-based models [1, 2] for our recurrence score dataset serve as a poignant reminder of the challenges of developing a one-size-fits-all solution [36]. Visualizations such as gradCAM [37], colormaps [38], and attention maps serve as informative tools, unveiling not only the model's class judgement but also the extent of fluctuations in the class judgment scores. In digital pathology, this insight is of paramount importance. However, even when we manage to acquire annotations with our GUI-based feedback collection system, uncertainties can persist. It's worth noting that despite the best efforts, mislabeling can have a detrimental impact on accuracy. Establishing precise biological delineations becomes particularly challenging in subtle cases where even experts encounter difficulties distinguishing between subtypes, such as *high*-risk vs. *low*-risk. We employ interpolation and smoothing techniques, such as thin spline interpolation, to mitigate such uncertainties. We construct an approximating function that captures crucial patterns within the data while minimizing noise. The standard diagnostic reference often originates from a single expert pathologist in our study, encompassing subjective interpretations. Consequently, it becomes more advantageous to incorporate a loss function that assigns weights to annotation uncertainty during the model re-training

process. We add a Mean Squared Error (MSE) loss term to the existing set of loss terms of the attention-based models during their re-training.

In this chapter, we explain the fundamental concept of Multiple Instance Learning (MIL), with particular emphasis on acquiring attention scores from WS attention-based models - both convolution and transformer architectures.

## 3.2 Extracting Attention scores

Standard benchmark natural image datasets such as ImageNet [25] encompass thousands of images with an average resolution of 469-by-387 pixels, commonly subjected to pre-processing by downsampling to 256-by-256 pixels to serve as inputs to Deep Learning architectures. Current advancements in hardware and software facilitate the parallelization of computations and efficient training of various WS attention-based models using such image dimensions. However, pathology presents a distinct challenge where the available training WSIs are limited (ranging from $10^1$ to $10^3$) and where each WSI comprises billions of pixels (approximately 100k-by-100k pixels). Furthermore, it is frequently the case that each WSI has a single label.

Given the training data, in contrast to typical scenarios for ImageNet [25], where one label corresponds to an individual image of an average resolution of 469x387 pixels, we now confront a situation with a single class label for a huge WSI (referred to as a **bag**) of multiple tiles or patches (**instances**). This notion called **Multiple Instance Learning**, was initially introduced by [39] as a solution to predict the activity of molecules in drug-related tasks.

Within WS networks, the task involves propagating the information represented by the bag label through the entirety of the network [40]. In conventional scenarios, one tries to find a mapping from an instance $x \in R^D$ to a label $y \in 0, 1$, whereas in MIL, the objective shifts to find a mapping from a bag of instances $X = x_1, ..., x_N$ to a label $Y \in 0, 1$. [26] highlights one possible assumption of MIL comprising the encoding and the aggregation stages as follows:

$$Y = \begin{cases} 0, & \text{iff } \sum_{k=1}^{N} y_k = 0, \\ 1, & \text{if any } y_k = 1 \end{cases} \tag{3.1}$$

Using Bernoulli distribution to model the probability of $Y$ given the bag of instances $X$:

$$p(Y|X) = S(X)^Y (1 - S(X))^{(1-Y)} \tag{3.2}$$

where, $S(X) = p(Y = 1|X)$ is a scoring function of a bag $X$ that is considered permutation invariant (a.k.a. a symmetric function) $iff$:

$$S(x_1, ..., x_N) = S(x_{\sigma(1)}, ..., x_{\sigma(N)}) \tag{3.3}$$

for any permutation $\sigma$.

For a set of instances $X$, $S(X) \in R$ is a symmetric function $iff$ it can be decomposed in the following form:

$$S(X) = g\left(\sum_{x \in X} f(x)\right) \tag{3.4}$$

where $f$ and $g$ are suitable transformations.

For any $\epsilon > 0$, a Hausdorff continuous symmetric function $S(X) \in R$ can be arbitrarily approximated by a function in the form $g(max_{x \in X} f(x))$, where $max$ is the element-wise vector maximum pooling function and $f$ and $g$ are continuous functions, that is,

$$|S(X) - g(max_{x \in X} f(x))| < \epsilon \tag{3.5}$$

Thus, given any $S(X)$, the procedure involves embedding all instances into a lower-dimensional space through $f$ and combining these embedded instances using a permutation-invariant function such as the summation or maximum in eq3.4, eq3.5. Mapping these embedded instances to a single scalar value, called score, uses $g$.

The collection of embeddings is denoted as $H = h_1, ..., h_N$, where each $h_k \in \mathbb{R}$. Typically, pooling layers in a network are employed to reduce the dimensions of the latent space after each layer of neurons. In the context of the MIL problem, these

pooling layers serve the additional purpose of pooling instance representations to derive bag representations, achieved either through combining instance scores (using the instance-based approach) or instance embeddings (via the embedded-based approach). Consequently, we will now look into the predominant MIL pooling functions:

- Max - for a vector $h \in R^k$:

$$h_{slide} = \max_{k=1,\ldots,N} \{h_k\}$$

- Mean - calculates an average embedding:

$$h_{slide} = \frac{1}{N} \sum_{k=1}^{N} h_k$$

Using such pooling functions, for instance, max pooling within MIL scenarios entails the selection of an instance with the highest predicted score for the positive class for the ultimate slide-level prediction by a model. However, it relies solely on the gradient signal from that individual instance within each WSI to adjust the learning parameters of the model. Hence, [26] proposed attention-based MIL pooling, characterized by a weighted sum operation as follows:

$$h_{slide} = \sum_{1}^{N} a_k h_k$$
$$a_k = \frac{\exp\left\{\mathbf{w}^\top \tanh\left(\mathbf{V}\right) \mathbf{h}_k^\top\right\}}{\sum_{j=1}^{N} \exp\left\{\mathbf{w}_\top \tanh\left(\mathbf{V}\right) \mathbf{h}_j^\top\right\}} \tag{3.6}$$

where, $w \in R^{L\times 1}$ and $V \in R^{L\times D}$ are learnable parameters with hidden dimension $L$.

The attention-based MIL pooling layer employs an auxiliary network comprising two fully connected layers. In the initial hidden layer, the output of hyperbolic tangent activation function $\dot{\tanh}()$ is symmetric, unlike sigmoid or ReLU. The subsequent layer utilizes the softmax nonlinearity to ensure that the attention weights sum to 1. This design renders the MIL pooling fully trainable and flexible, with a straightforward interpretation of the attention weights.

### 3.2.1 Extracting attention scores from Convolution Architecture

In a general multi-class classification task, each instance (patch) has an unknown label $y_i \in C$ (instance-level) and an available $y \in C$ (slide-level) for a bag (WSI). Unlike the above binary pooling [26], [1] involves predicting distinct attention scores corresponding to the different categories. The attention-based pooling function aggregates slide-level representations from the patch-level representations for each category. One possible shortcoming of the attention-based MIL pooling layer is the $\tanh(\cdot)$ nonlinearity. The expressiveness of $\tanh(\cdot)$ is limited since it is approximately linear for $x \in$ [-1, 1] [26]. Hence, in Gated Attention (GA) in [1], patch-level feature $H$, attention score of the $k^{th}$ patch for the $m^{th}$ class, denoted $a_{k,m}$, is given by

$$\begin{aligned} a_{k,m} &= \frac{\exp\left\{\mathbf{W}_{a,m}\left(\tanh\left(\mathbf{V}_a\mathbf{h}_k^\top\right) \odot \text{sigm}\left(\mathbf{U}_a\mathbf{h}_k^\top\right)\right)\right\}}{\sum_{j=1}^{N}\exp\left\{\mathbf{W}_{a,m}\left(\tanh\left(\mathbf{V}_a\mathbf{h}_j^\top\right) \odot \text{sigm}\left(\mathbf{U}_a\mathbf{h}_j^\top\right)\right)\right\}} \\ \mathbf{h}_{slide,m} &= \sum_{k=1}^{N} a_{k,m}\mathbf{h}_k \end{aligned} \tag{3.7}$$

and WSI-level representation for the $m^{th}$ class is

$$W_{a,1}, ..., W_{a,k} \in R_{L\times 1} \tag{3.8}$$

where, $V \in R^{L\times D}$ and $U \in R^{L\times D}$ are learnable parameters with hidden dimension $L$.

### 3.2.2 Extracting attention scores from Transformer Architecture

It becomes essential to consider the dependencies between instances, which convolutions fail to address adequately. To overcome this i.i.d limitation, transformers, capturing long-range dependencies, prove promising [29, 41]. Within their self-attention

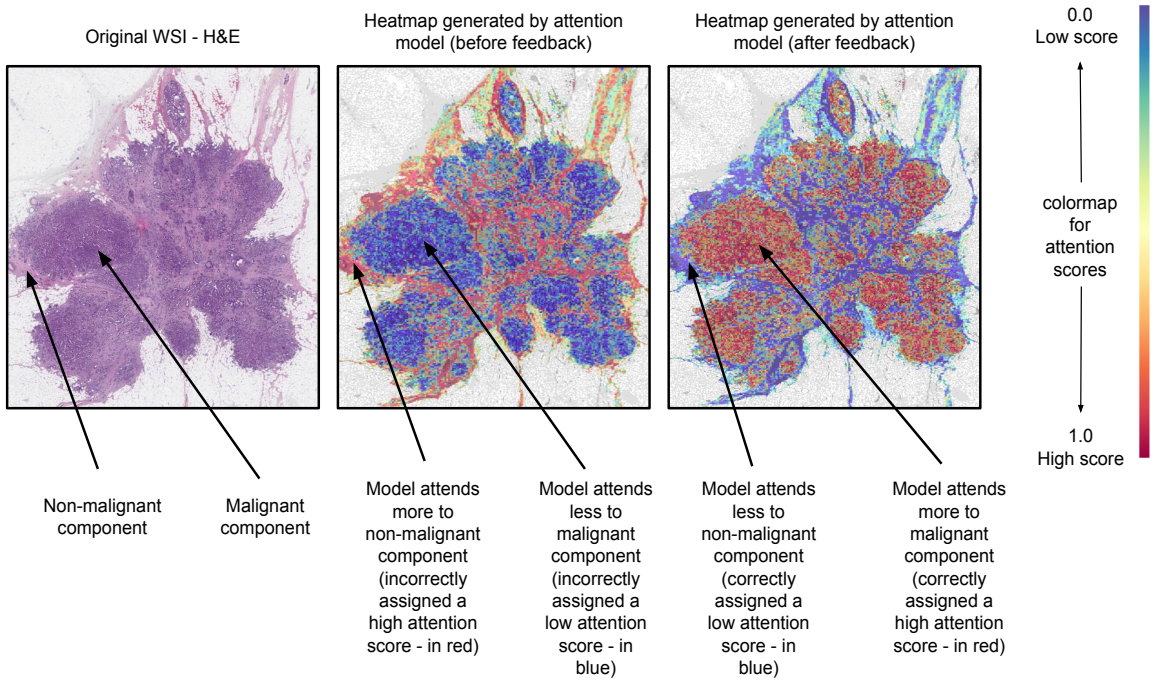



Figure 3.2: Visualization of heatmap based on the attention scores generated by the model before and after the feedback mechanism

block, an input sequence of $k$ tokens with dimensions $D$ corresponding to the instance features space $H$, project and extract feature representations: $W_Q \in R^{d\times d_q}$, $W_K \in R^{d\times d_k}$ and $W_V \in R^{d\times d_v}$. $Q$, $K$ and $V$ are as query, key and value, where $Q = XW_Q$, $K = XW_K$ and $V = XW_V$. The approximated self-attention form in [2] is as follows:

$$\hat{\mathbf{S}} = \text{softmax}\left(\frac{\mathbf{Q}\tilde{\mathbf{K}}^T}{\sqrt{d_q}}\right)\left(\text{softmax}\left(\frac{\tilde{\mathbf{Q}}\tilde{\mathbf{K}}^T}{\sqrt{d_q}}\right)\right)^{+}\text{softmax}\left(\frac{\tilde{\mathbf{Q}}\mathbf{K}^T}{\sqrt{d_q}}\right) \tag{3.9}$$

where $\tilde{\mathbf{Q}}$ and $\tilde{\mathbf{K}}$ is the selected landmarks from the original sequence of $Q$ and $K$, and $A\pm$ is a Moore-Penrose pseudoinverse of $A$ [2]. By doing this, the module with approximation processing can satisfy our case where the bag contains thousands of WSI patches.

### 3.2.3 Generating Heatmaps for WSIs

We capture the relevant attention information of each WSI from both model types for multiclass classification. In the convolution, we extract the patch attention scores from the multiple attention branches of the network and choose the one that corresponds with the ground truth label of the WSI. On the other hand, in the transformer architecture, we compute the average attention score over all the heads in a Multi-Head Self-Attention network for a patch.

To interpret the importance of different regions in a WSI, we normalize the above-extracted attention scores and scale them between 0 and 1.0, with 0.0 representing the lowest attention and 1.0 denoting the highest. We store the patch scores with their top-left coordinate information to visualize and interpret regions of malignancy (features characteristic of the class) and non-malignance (irrelevant features) using a diverging colormap ranging from red (1.0) to blue (0.0). An example of the heatmap visualization is shown in **Fig. 3.2**.

## 3.3 Judiciously engaging Pathologist's expertise

We develop a Graphical User Interface (GUI) for pathologists to facilitate the feedback collection for the above heatmaps based on their expertise and domain knowledge. By incorporating these annotations into the model's subsequent re-training process, the pathologist's insights and interpretations become an intrinsic component of the model's learning trajectory. This integration aims to harmonize the model's attention scores with the pathologist's intuitive grasp of malignant features in the WS and capture crucial diagnostic patterns aligning closely with established clinical standards. In the subsequent chapter, we explain the GUI's design and the processing of pathologist's annotations.

# Chapter 4

# User-friendly feedback collection system: *The Graphical User Interface (GUI)*

## 4.1 Design Principles

This chapter introduces an interactive and user-centric feedback acquisition system tailored for annotating heatmaps produced by attention-based models. Our approach encompasses the development of a customized Graphical User Interface (GUI), which integrates Openseadragon and Openslide technologies. This GUI provides pathologists with a powerful capability to directly delineate free-form bounding boxes on the heatmaps of the WSI, effectively eliminating the need for time-consuming manual annotations. The schematic representation of our feedback collection system in **Fig. 4.1** streamlines the process of incorporating expert insights into the model's learning process to enhance its overall performance.

### 4.1.1 GUI foundational requirements

When devising the GUI, we consider the fundamental prerequisites that this tool should satisfy:

- Handle large high-resolution images - As elucidated in **Section 2.2**, digital slides pan across gigapixel resolution supported with a *tiled pyramid* format

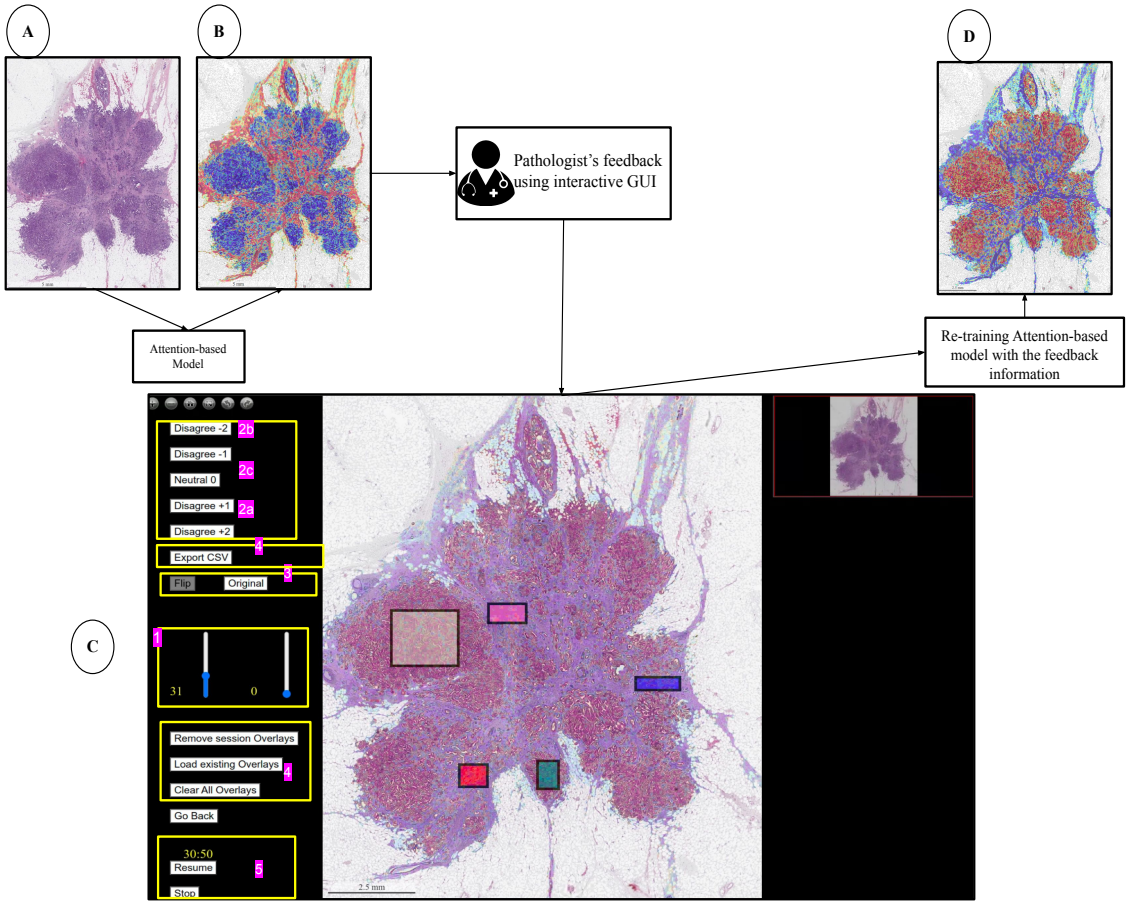



Figure 4.1: **A**- the original WSI; **B**- initial heatmap generated by the attention-based model; **C**- interactive GUI for providing the feedback; **D**- heatmap of the WSI after incorporating the feedback
In **C**, notice that the GUI user has opted to reverse the heatmap using **Flip** button for annotation. This deliberate adjustment of altering the attention scores results in a heatmap in **D** that meets the desired criteria in **C**.
**NOTE:** The bounding boxes in **C** are drawn by the author, who is not a clinical expert. Different coloured boxes are for illustrative purposes only.

when displayed on a monitor screen. Hence, the GUI can zoom, pane, and navigate, closely mirroring how pathologists utilize a microscope to scrutinize different regions in the slide.

- Overlays - Beyond the original slide, a crucial requirement is to display heatmaps and annotations from the pathologist with coherence. When superimposed upon the original image, the heatmaps and annotations should remain accessible to the pathologist, enabling them to explore the tissue beneath. Additionally, it should offer the flexibility to adjust the opacity of overlaid heatmaps.

- Annotation support - The viewer should allow pathologists to generate annotations over regions of their choosing, facilitated by appropriate event handlers.

- Interactiveness - Pathologists who frequently use the tool must comprehend its functionalities and be content with its performance. For example, the speed of zooming and panning is vital, as noticeable delays can disrupt their workflow.

Openseadragon [42] is pivotal for rendering high-resolution images, thereby preserving the intrinsic visual quality inherent in WSIs. Complementing this functionality, Openslide [43] facilitates access to the content of WSIs, ensuring the smooth operation of the GUI even in the presence of these intricate and high-dimensional images.

### 4.1.2 Design Specifications

OpenSeadragon [42] is an open-source web viewer constructed using JavaScript, designed to accommodate high-resolution images with zooming capabilities. This viewer exhibits compatibility with various server protocols, allowing integration with zoomable image formats, including DZI, TMS, Zoomify, and IIF. Its primary role is to enable web-based visualization, ensuring smooth navigation through exceedingly large files without discernible latency. It complements zooming and panning with optional

controls like toolbars, rotation features, overlays, and multi-image support. Additionally, it supports plugins, which can extend its capabilities further, encompassing annotations, magnification, navigation aids, and even the capture of screenshots. OpenSeadragon, by default, works with DZI images and relatively simple formats like PNG and JPEG. It accommodates various image-serving protocols, including Openslide's DeepZoom, which facilitates zooming for *tiled pyramids*.

Deep Zoom [43] serves images in a multi-resolution manner, allowing loading and panning operations. During the initial load, a low-resolution version of the image is presented, progressively transitioning to higher resolutions as they become available. This mechanism underpins the gradual shift from a blurry-to-sharp experience. OpenSlide Python [43] extends its capabilities to generate individual Deep Zoom tiles from slide objects. This feature is valuable for exhibiting WSIs within a web browser, all without necessitating the complete conversion of the entire slide into the Deep Zoom format.

The integration of Flask [44], a lightweight web framework, in combination with Openseadragon and Openslide, serves as a backbone for constructing the web GUI, optimizing the workflow of annotating heatmaps generated by attention models. This integration establishes a dynamic and responsive context that facilitates interaction between pathologists and GUI.

## 4.2 GUI Functionality

In addition to the above inherent features from its constituents - Openseadragon and Openslide, our GUI offers the following tailored features for interactive feedback from pathologists.

1. **The Flip button** - Given the considerable contradiction observed in the heatmaps (in **Chapters 1, 2, 3**) due to the *negative picture* of attention, it would be laborious for pathologists to draw bounding boxes for every region of contradiction

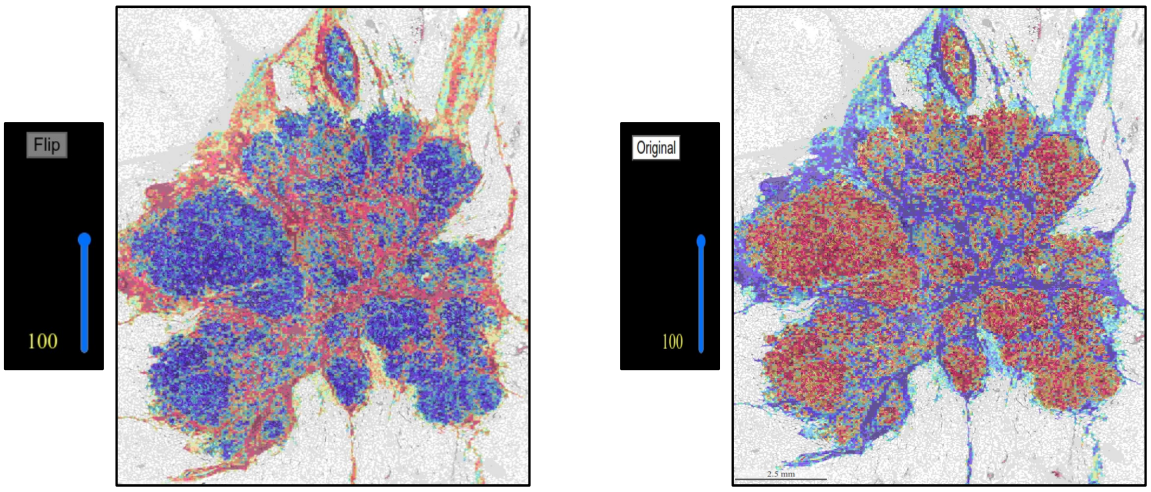



Figure 4.2: The Flip button

in a WSI (here, drawing bounding boxes on every blue zone to increase it's attention score). To address this, we introduce a *Flip* button in our GUI to streamline the process. It switches between two versions of the heatmap - the *original* attention scores generated by the model and their corresponding *reverse* values(**Fig. 4.1C**; zoomed in: **Fig. 4.2**- greyed out indicates the user has chosen the reverted heatmap).

For instance, if the attention for a patch is 12%, its reverse value would be 88%. This feature empowers pathologists to quickly switch between the model's prediction (*Original* button) and its opposite, saving them from the exhaustive task of manually annotating the entire WSI when a complete disagreement arises.

2. **Dual Heatmap Opacity** - We provide an opacity bar to adjust the transparency of both versions of the heatmap overlaid on the WSI.

3. **Annotating Options** - We provide distinct modes or options, each featuring specific functionalities related to the bounding boxes. These bounding boxes can vary in size, each serving a particular function:

   (a) *Increasing Attention Scores* - To enable pathologists to enhance attention

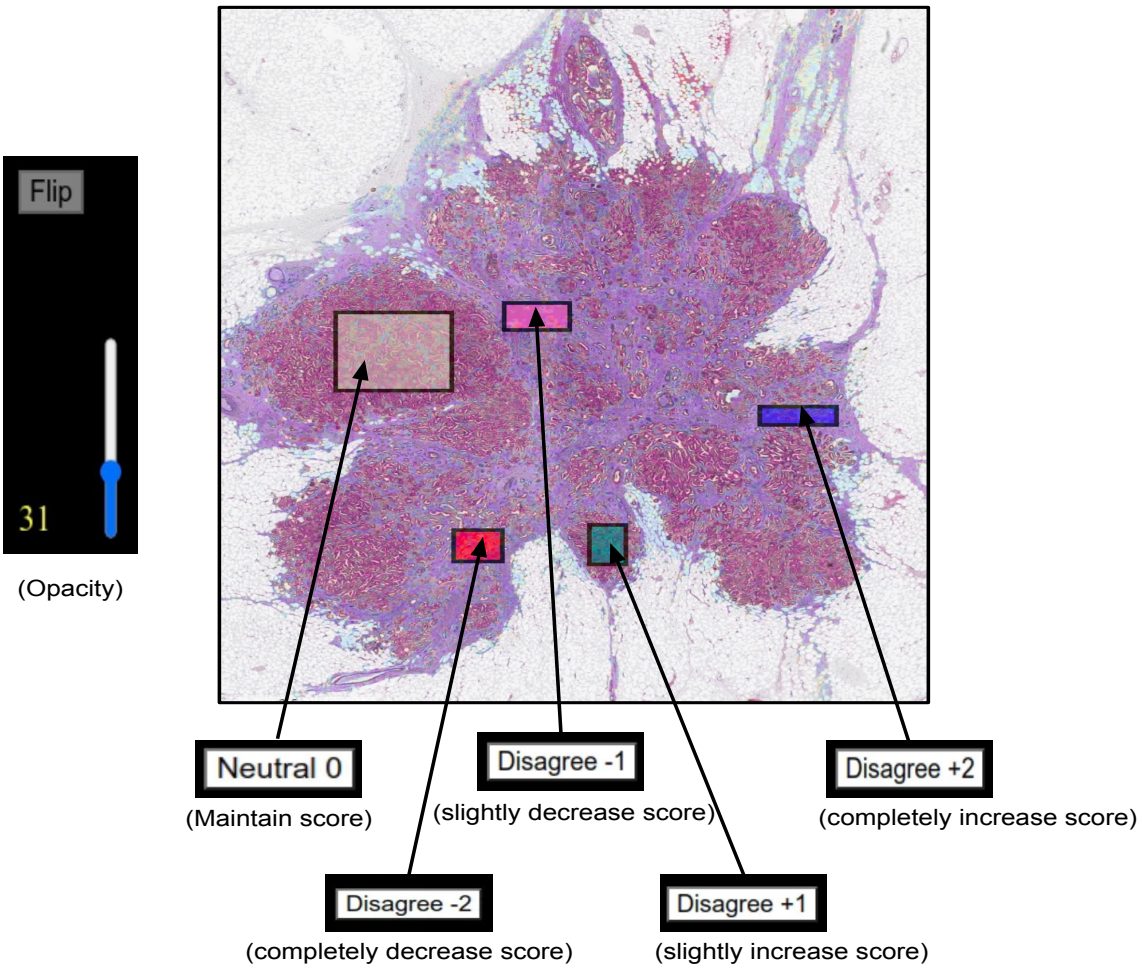



Figure 4.3: Annotating options

scores where necessary, we present two dedicated buttons - *Disagree +1* and *Disagree +2*.

Pathologists can incrementally raise the attention scores assigned to particular regions. The *+1* button corresponds to a minor increment, while the *+2* button signifies a more substantial enhancement. In the visualization depicted in **Fig. 4.1C** (zoomed in: **Fig. 4.3**), bounding boxes coloured in green and blue correspond to the *Disagree +1* and *Disagree +2* buttons, respectively.

(b) *Decreasing Attention Scores*: To allow pathologists to reduce attention scores for specific regions, we provide corresponding buttons - *Disagree -1* and *Disagree -2*.

Pathologists can systematically lower the attention scores assigned to particular regions. The *-1* button denotes a slight decrement, while the *-2* button represents a more pronounced reduction. In **Fig. 4.1C** (zoomed in: **Fig. 4.3**), bounding boxes in pink and red align with the *Disagree -1* and *Disagree -2* buttons, respectively.

(c) *Maintaining Correct Scores*: Acknowledging instances where the model's attention scores are accurate, we introduce the *Neutral 0* button. Pathologists indicate they agree with the model's predictions to maintain the existing attention scores for the designated regions using this functionality (transparent bounding box in **Fig. 4.1C** (zoomed in: **Fig. 4.3**)).

4. **Export to CSV** - Pathologists can export their annotations into a Comma-Separated Values (CSV) file. This feature empowers pathologists to maintain a tangible record of their annotations. Moreover, the GUI facilitates viewing or hiding these annotations through dedicated buttons such as *Load Existing Overlays* and *Clear All Overlays* in **Fig. 4.1C**.

In scenarios where errors or inaccuracies may arise during an ongoing session,

the *Remove Sessions Overlay* button becomes a valuable tool. Pathologists can swiftly eliminate all annotations associated with the current session. This reset mechanism empowers pathologists to rectify any discrepancies and start afresh, ensuring the provision of precise and reliable annotations.

5. **Stopwatch** - Pathologists can track the time invested in annotating a WSI. Should interruptions or distractions occur during the annotation process, pathologists can conveniently pause the stopwatch and resume the process without compromising workflow continuity. This feature is valuable in accommodating the dynamic nature of pathologists' work environments (in **Fig. 4.1C**).

   The stopwatch also caters to scenarios where pathologists might prefer to export the ongoing session's CSV file, pause the annotation process, and resume annotations from the same point. This fluidity is facilitated through the GUI's *Load Existing Overlays* button to retrieve the timestamp of the last saved bounding box. This feature ensures a seamless annotation process and the elapsed time accurately reflects the cumulative effort invested in the task.

Our GUI enriches the interaction between pathologists and the heatmaps generated by attention-based models by providing tools to streamline annotation management, time tracking, and error rectification. Through export alternatives, overlay controls, and stopwatch functionalities, our GUI ensures a holistic and user-oriented mechanism for collecting feedback.

## 4.3 Processing pathologist's annotation

### 4.3.1 Patch Co-ordinates mapping

After the feedback collection process, we extract information from the CSV files. These CSV files constitute structured tabular datasets characterized by rows and columns. Pertinent columns relevant to processing the pathologist's annotations include:

1. **Flip** - This column denotes the status of the flip button, with *0* representing *Original* and *1* representing *Flip*. Pathologists can toggle between viewing and assessing the *Original* or *Reverse* heatmap versions and fine-tune the heatmap based on their annotation preferences to increase, decrease, or maintain the existing scores.

2. **Mode** - The mode column signifies the bounding box type based on the pathologist's choice and records its corresponding numerical value from the set *-2, -1, 0, 1, 2*.

3. **Top-left Coordinates** - This pivotal data component enables the precise localization of the annotated region. The *(x, y)* coordinates of the top-left corner of the bounding box with the *height* and *width* of the bounding box pinpoints the patches that fall within the designated bounding box area.

4. **Time** - The cumulative time taken by the pathologist to draw each bounding box on the WSI encompasses the entirety of the process, irrespective of pauses or multiple annotation sessions. The time logged for a particular WSI is the cumulative duration from the first bounding box.

Some WS attention-based models for WSIs face resource limitations during training, which prevent them from performing on-the-fly feature extraction. To address this challenge, these models partition the WSI into smaller patches $X = x_1, ..., x_N$, such as 256-by-256 pixels (typical MIL). They then record the top-left coordinates of these patches and conduct feature extraction on each patch, storing the resulting features locally. In our approach, we leverage this stored patch information to identify the top-left coordinates and perform exhaustive mapping to search for the nearest patch coordinate to each corner of the bounding box. For instance, if we consider extracting coordinates under the green bounding box as depicted in **Fig. 4.4**, the patches 'X' will be excluded and the patches '+' will correspond to the ones with the

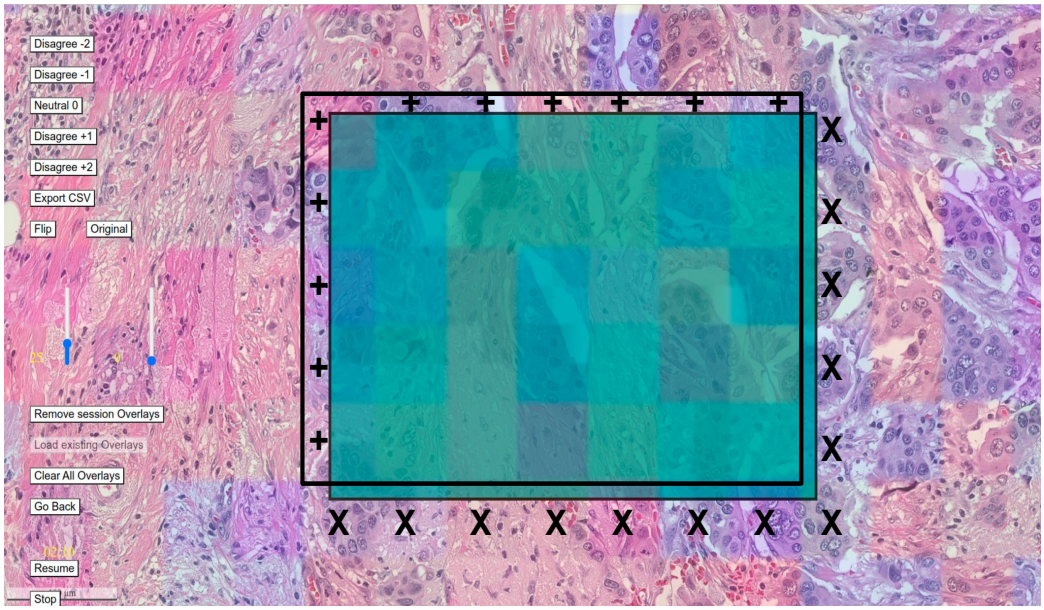



Figure 4.4: Mapping patch coordinates

nearest coordinates in the vicinity. This process hinges on identifying the shortest linear distance between the top-left coordinate of the patch and the desired bounding box corner. Through this, we associate the corresponding annotations provided by pathologists, where applicable. It's important to note that patches without feedback are retained and not discarded.

### 4.3.2 Thin spline interpolation: *Expected scores*

Following the identification of specific patches within individual bounding boxes, along with the extraction of essential data like the top-left patch coordinate details and the assigned box mode, we incorporate the concept of thin spline interpolation (as shown in **Fig. 4.5**). It constitutes a vital phase to refine and approximate the relationship between bounding box modes and normalized attention scores. We select this interpolation because of its proficiency in generating a continuous and coherent curve that adeptly navigates through the provided data points.

The continuous curve represents the underlying trends that reside within the data. In contrast to linear interpolation, which merely constructs straight lines connecting

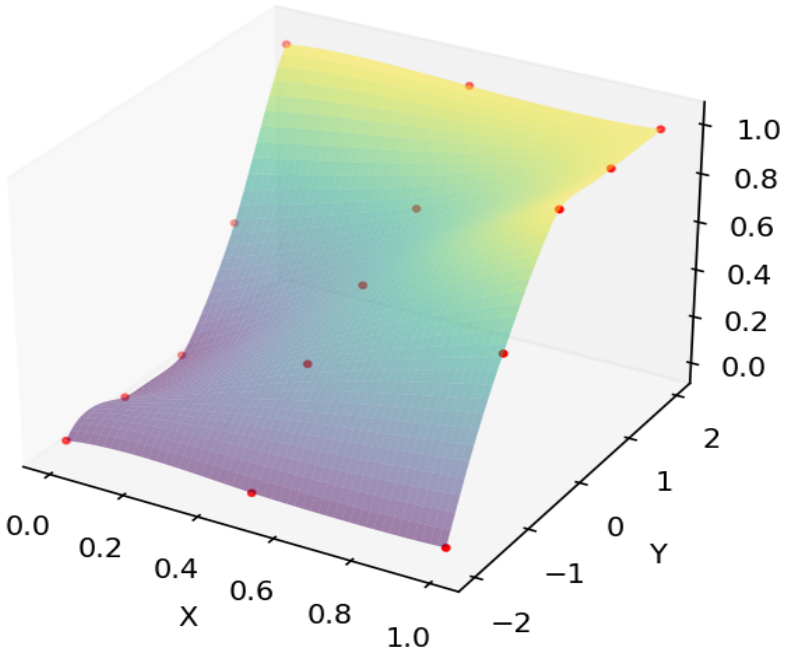



Figure 4.5: Thin Spline Interpolation

individual data points, the adoption of thin spline interpolation results in a curved trajectory. It mitigates the potential distortion introduced by noisy or outlier data that might otherwise not conform perfectly to the attention scores.

$z = f(x, y)$;

$x = [0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0.5, 0.5, 0.5, 0.5, 0.5]$,

$y = [0, -1, -2, 1, 2, 0, 1, 2, -1, -2, 0, -1, -2, 1, 2]$,

$z = [0, 0, 0, 0.4, 1, 1, 1, 1, 0.6, 0, 0.5, 0.35, 0, 0.65, 1]$;

where, $x$ = Normalized attention scores, $y$ = bounding box modes and $z$ = interpolated scores.

The interpolation is clipped between the values 0 and 1 to ensure coherence. Following this process, we store the patch coordinates, the newly generated attention scores, now called the *expected scores*, and their associated metadata in a CSV file. The interpolation bridges the annotations and attention scores, translating expert insights into a refined representation, the *expected scores* that guide the model's learning process.

## 4.4 Additional Loss term: Sum of Squared Errors (MSE)

We incorporate the pathologist's feedback for every WSI into the model's learning process through a least squares error correction (MSE loss) to penalize the difference between the **expected scores** and the model's generated attention scores during the learning process, now called as the **current scores**. While alternative loss functions could be employed, the simplicity of the MSE loss is suitable for continuous and real-valued attention scores. It provides a well-understood and interpretable metric for quantifying the alignment between current scores and the pathologist's refined expectations. The objective function is as follows:

$L_{feedback} = MSE(input = x_{new}, target = z)$

$$L_{total} = L_{model} + \lambda L_{feedback} \tag{4.1}$$

where,

- $L_{model}$ represents existing loss function(s) in any attention-based model;
- $L_{feedback}$ represents the MSE loss for feedback;
- $\lambda$ represents the coefficient of the MSE term;
- $L_{total}$ represents the cumulative or gross loss for the attention-based model.

The coefficient $\lambda$, acting as a weighting factor, regulates the influence of the MSE loss within the overall loss function. Its role is pivotal in determining the balance between the model's original loss function(s) $L_{model}$ and the extra MSE loss $L_{feedback}$. Through $\lambda$, we can modulate the emphasis placed on correcting the attention scores based on pathologist feedback and controlling the extent to which the model adapts to the feedback, ensuring that the incorporation of pathologist insights balances with the model's inherent learning objectives. Hence, the choice of MSE, the coefficient

$\lambda$, and the incorporation into the cumulative loss $L_{total}$ create a dynamic and balanced learning framework that fuses automated predictive capabilities and human expert insights for more accurate and clinically meaningful attention-based model predictions.

# Chapter 5

# Experiments and Results

We conduct a series of trials to assess the efficacy of attention-based models within the context of our research inquiries. Firstly, we present an overview of our in-house dataset and elucidate the parameter settings for training attention-based networks. Then, we outline the evaluation metrics employed to gauge the predictive ability of the models.

Through these experiments, our primary objective is to shed light on the impact of incorporating human feedback in the model training process and demonstrate its benefits in enhancing the performance of attention-based models and generating better representative heatmaps.

## 5.1 Experimental setup

### 5.1.1 Dataset

Our study utilizes an in-house dataset comprising 727 Whole Slide Images (WSIs) from Health Centres in Alberta (approved by the Health Research Ethics Board of Alberta (HREBA.CC-19-0347)) and the Department of Pathology and Laboratory Medicine, Dalhousie University, Halifax, Canada. These WSIs underwent H&E staining and are digitized via the GT450 Aperio scanner, capturing high-resolution images at a 40x objective magnification.

Within this dataset, each WSI is accompanied by pertinent clinical metadata and

results of routinely performed pathological assessments for BC, including Estrogen Receptor (ER), Progesterone Receptor (PR), in addition to the ground truth prediction RS, offering a comprehensive contextual framework.

To facilitate dataset organization and classification, we partition the samples into two categories based on their RS scores (OncotypeDX®), spanning from 0 to 100. WSIs with RS values $\leq 25$ are classified as **low**, while those exceeding 25 are labelled as **high**, adhering to the classification established by the TAILORx study as mentioned in **Section 2.1**. To ensure the statistical robustness of our evaluation, we adopt a 10-fold Monte Carlo cross-validation strategy from [1]. This approach involves the division of the dataset into training, validation, and test subsets, maintaining a proportion of 70:15:15 for the 10-fold cross-validation, respectively.

To effectively manage the storage and organization of the digitized slide raw files, we employ a 10TB hard drive. This choice of storage medium reflects a careful consideration of the substantial data volume generated by the high-resolution digitization process. By utilizing this storage solution, we ensure the accessibility and integrity of the dataset, enabling seamless retrieval and manipulation for subsequent analysis and experimentation.

### 5.1.2 Training details

We select two attention-based models for our study: the convolution-based **CLAM** (Clustering-constrained Attention Multiple Instance Learning) [1] and **TransMIL** (Transformer based Correlated Multiple Instance Learning) [2]. We begin by loading WSIs into the system's memory. The authors of TransMIL have stated that they have incorporated the pre-processing steps of CLAM. Hence, we follow the *Pytorch* implementation of CLAM for creating patches and extracting features for both CLAM and TransMIL models. We convert the images to the HSV colour space and generate binary masks for tissue regions by thresholding the saturation channel. Subsequently, we extract patches measuring $256 \times 256$ pixels from the delineated

foreground contours and record their corresponding top-left coordinates. We employ the ImageNet [25] pre-trained ResNet50 model for converting each extracted patch into a 1024-dimensional feature vector representation. This transformation enhances the capacity of the patches processed during the training of CLAM and TransMIL models.

Our experimentation is conducted on a workstation equipped with an NVIDIA GeForce 3090 GPU, offering the computational power required for efficient execution. By adhering to these steps and leveraging the computational resources at hand, we explore the performance of both CLAM and TransMIL in our study.

## 5.2 Standard reference: The Universal model-independent representation

We initially generate heatmaps using the vanilla versions of CLAM [1] and TransMIL [2], focusing on the optimal fold within the 10-fold Monte Carlo cross-validation setup. Our feedback process centers around the heatmaps generated solely by CLAM. After the expert meticulously corrects a WSI heatmap, this representation works as a diagnostic attention map according to the clinician's standards.

In other words, the heatmap derived from any vanilla attention model, coupled with the subsequent expert corrections, forms the standard reference of that WSI. This reference portrays a universal model-independent representation of WSI's diagnostic expectations. It means that the corrected heatmap pinpoints regions of pronounced diagnostic importance within the WSI (as depicted in **Fig. 5.1D**).

Throughout the feedback process, we uphold the anonymity of the WSI, including classification probabilities, patient identification, and RS of the slide. This anonymity for the pathologist mitigates any upfront bias for a tumour grade or RS. The attention-based models receive WSIs without molecular information except for weak labels, and the pathologist is blind to such details, similar to the models. On average, our pathologist took approximately *34* to *52* seconds to annotate a single WSI. We employ

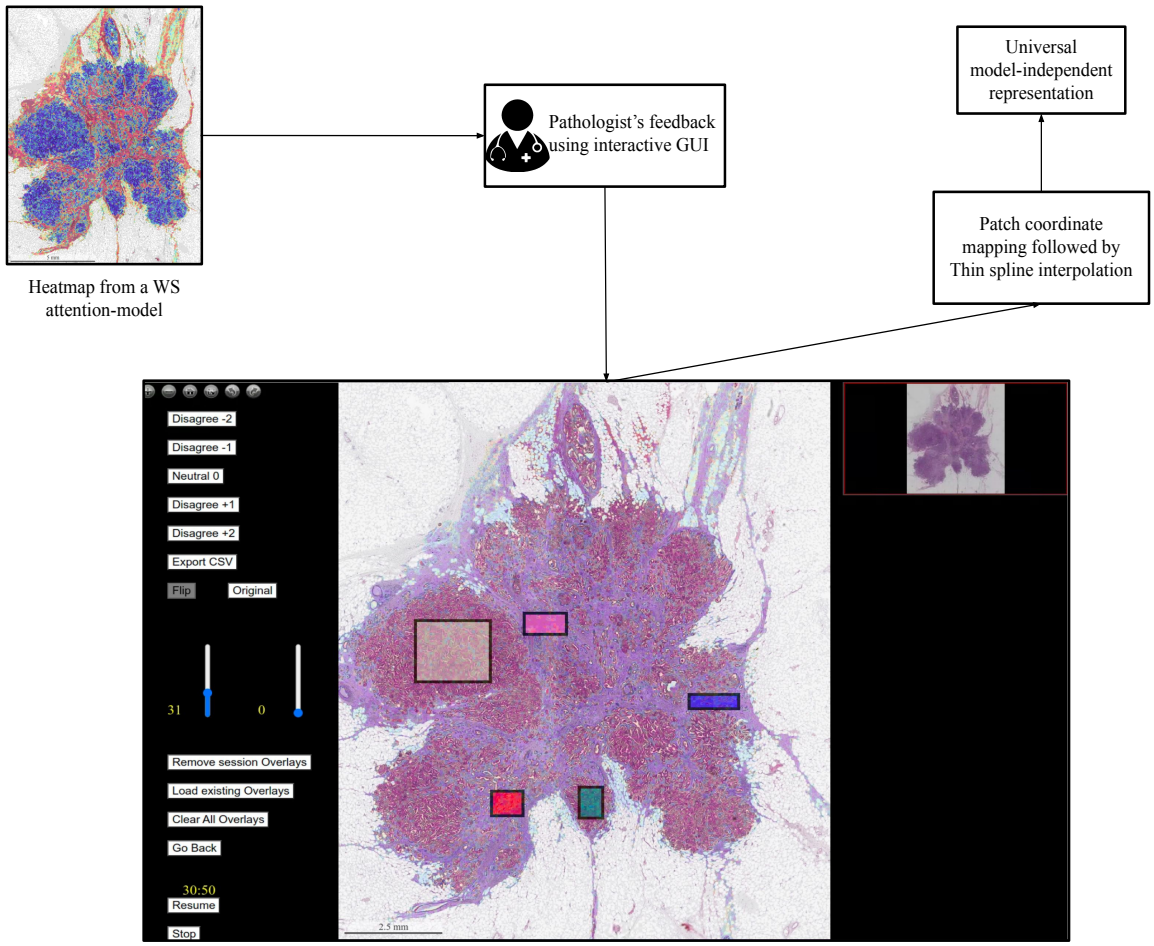



Figure 5.1: The Universal model-independent representation
**NOTE:** The bounding boxes are drawn by the author, who is not a clinical expert. Different coloured boxes are for illustrative purposes only.

thin spline interpolation, as mentioned in Chapter **4**, on each WSI's annotations to facilitate the incorporation of the expert annotations.

## 5.3 Performance of our Approach

### 5.3.1 Hyperparameter optimization using Optuna framework

After successfully establishing a standard reference for the WSIs through expert annotations, our focus shifts to re-training the CLAM [1] and TransMIL [2] models. This process integrates the difference between *expected scores* and *current scores* for WSIs as MSE loss term as described in Chapter **4**. This aggregated loss, a combination of the existing loss function(s) and the MSE loss is then weighted and minimized during the training phase.

$$L_{total} = L_{model} + \lambda L_{feedback} \tag{5.1}$$

where

- $L_{model}$ represents existing loss function(s) in any attention-based model;
- $L_{feedback}$ represents the MSE loss for feedback;
- $\lambda$ represents the coefficient of the MSE term;
- $L_{total}$ represents the cumulative or gross loss for the attention-based model.

For the optimization of hyperparameters aimed at minimizing aggregate loss and maximizing model performance on the validation dataset, we leverage the Optuna framework [45]. Hyperparameter optimization determines the optimal parameters that govern the learning algorithm's performance. In our case, this involves parameters - learning rates and coefficient $\lambda$. Optuna provides a systematic search space exploration, leading to the identification of the best-performing hyperparameter combination. Specifically, we define the search space for **CLAM**, including **learning**

**rates (2e-3, 2e-4, 2e-5, 2e-6)** and the coefficient $\lambda$ **(0.5, 5, 50, 500)**. For **TransMIL**, the search space incorporates **learning rates (2e-4, 2e-5, 2e-6, 2e-7)** and $\lambda$ **(5, 10, 20, 50, 100)**.

The Optuna search process took 20 GPU days for CLAM and 27 GPU days for TransMIL to explore all combinations within the defined search space. Optuna's outcomes provide the optimal hyperparameter combinations for each dataset combination in the 10-fold cross-validation. In particular, for *CLAM*, the *optimal learning rate of **2e-4*** and the $\lambda$ *of **50*** proved the most effective. For *TransMIL*, the *best learning rate was **2e-5***, and ***all*** $\lambda$ *values performed well.*

### 5.3.2 Results analysis and comparison

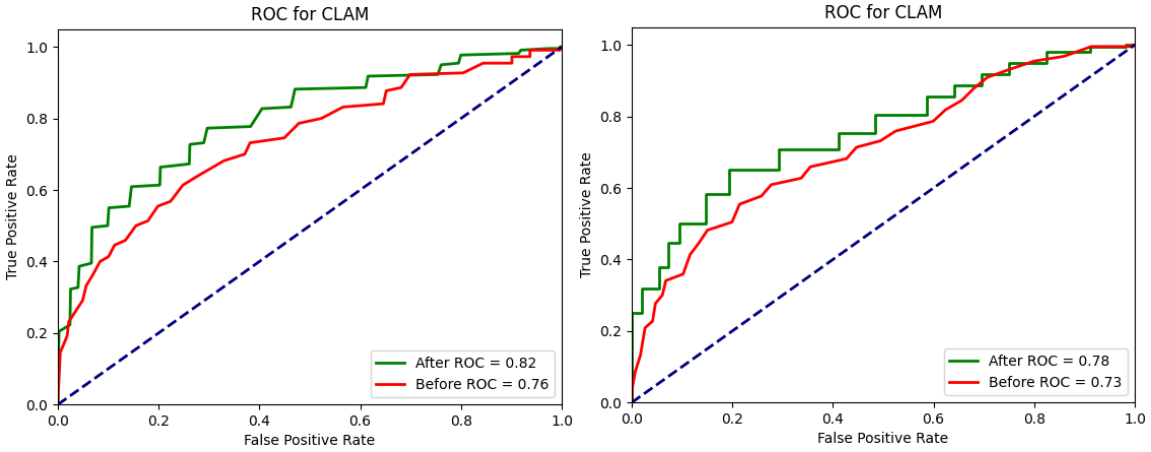
After conducting hyperparameter optimization, it becomes evident that each fold displays a peak performance when subjected to a specific hyperparameter configuration. [1] and [2] assess the performance of the models through classification metrics - Accuracy and Area Under the Curve (AUC) [1]. For instance, in the case of CLAM, the sixth fold performs well when utilizing a *learning rate* of *2e-4* and coefficient $\lambda$ of *50*, whereas, for the first fold, a *learning rate* of *2e-4* and coefficient $\lambda$ of *5* is the best. Mirroring the practice observed in [1], instead of relying on a single, top-performing fold, we gauge the efficacy of our approach by considering the average performance and the standard deviation across all ten folds for both CLAM [1] and TransMIL [2] depicted as *CLAM After* and *TransMIL After*, respectively in **Table 5.1**.

**Fig. 5.2** and **Fig. 5.3** shows the ROC AUC curves for both CLAM and TransMIL models following the incorporation of expert feedback. Each curve represents the trade-off between the True Positive Rate (sensitivity) and the False Positive Rate (1-specificity) at various classification thresholds using *sklearn* library. It's important to note that the ROC curves showcase the models' performance across a range of

[1]NOTE: The extra MSE loss magnitude is on the order of 1e-2

Table 5.1: Prediction metrics for CLAM and TransMIL before and after the pathologist's feedback

| Model | Feedback | Test AUC | Val AUC | Test Acc | Val Acc | Cosine Similarity |
|---|---|---|---|---|---|---|
| CLAM | After | 0.781±0.055 | 0.819±0.050 | 0.824±0.046 | 0.839±0.031 | 0.861 |
| | Before | 0.731±0.064 | 0.756±0.058 | 0.786±0.056 | 0.8012±0.039 | 0.810 |
| TransMIL | After | 0.813±0.066 | 0.8344±0.049 | 0.8463±0.032 | 0.8554±0.015 | 0.866 |
| | Before | 0.778±0.055 | 0.8191±0.053 | 0.7718± 0.029 | 0.8243±0.036 | 0.769 |

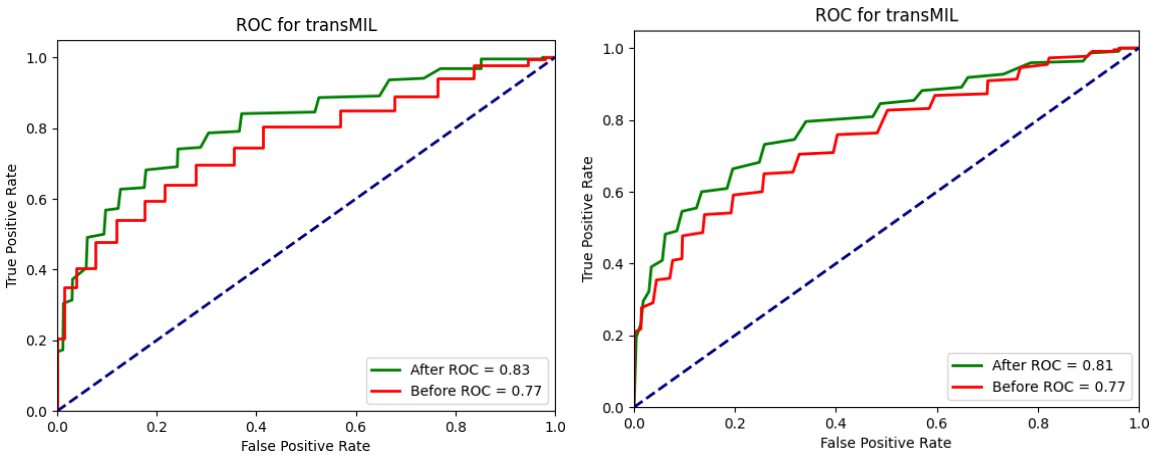



(a) Validation AUC for CLAM (b) Test AUC for CLAM

Figure 5.2: Validation and test AUC for CLAM

threshold values, providing a view of their discriminative abilities. Specifically, we examine the models' sensitivity and specificity across a range of threshold values to assess their robustness and effectiveness in distinguishing between *high* and *low* for the slides, with higher AUC scores indicating better discriminative performance. We observe a significant 5% increase in validation-test AUC and a 4% increase in validation-test accuracy on average for CLAM. For TransMIL, there is an average increase of 4.5% in validation-test AUC and a 3% increase in validation-test accuracy.

In our study, we utilize cosine similarity to quantify the degree of similarity or agreement between the attention scores generated by CLAM / TransMIL and the expected scores. Cosine similarity is a mathematical measure used to determine the




(a) Validation AUC for TransMIL (b) Test AUC for TransMIL

Figure 5.3: Validation and test AUC for TransMIL

cosine of the angle between two vectors in a multi-dimensional space.

$$\text{Cosine Similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

A cosine similarity score of 1 indicates that the vectors are perfectly aligned, meaning that the attention scores by CLAM / TransMIL and the pathologist's GUI-based expected scores are in complete agreement. On the other hand, a score of 0 implies that the vectors are orthogonal or unrelated, signifying no agreement between the attention scores. A score of -1 indicates complete misalignment or dissimilarity between the scores. We notice an improvement in cosine similarity between the pathologist's GUI-based attention scores and trained models' attention maps after feedback - 5% and 10% increase for CLAM and TransMIL, respectively.

To determine whether there are statistically significant differences in the cosine similarity results between attention scores generated by CLAM / TransMIL and the expected score, we perform the Wilcoxon Signed-Rank Test. Specifically, we are interested in comparing the similarity between attention scores generated by these models before and after feedback interventions. In **Fig. 5.4**, we notice a low p-value ($p < 0.001$) indicating strong evidence of a significant difference in the observations. We reject the null hypothesis because there is a statistically meaningful difference

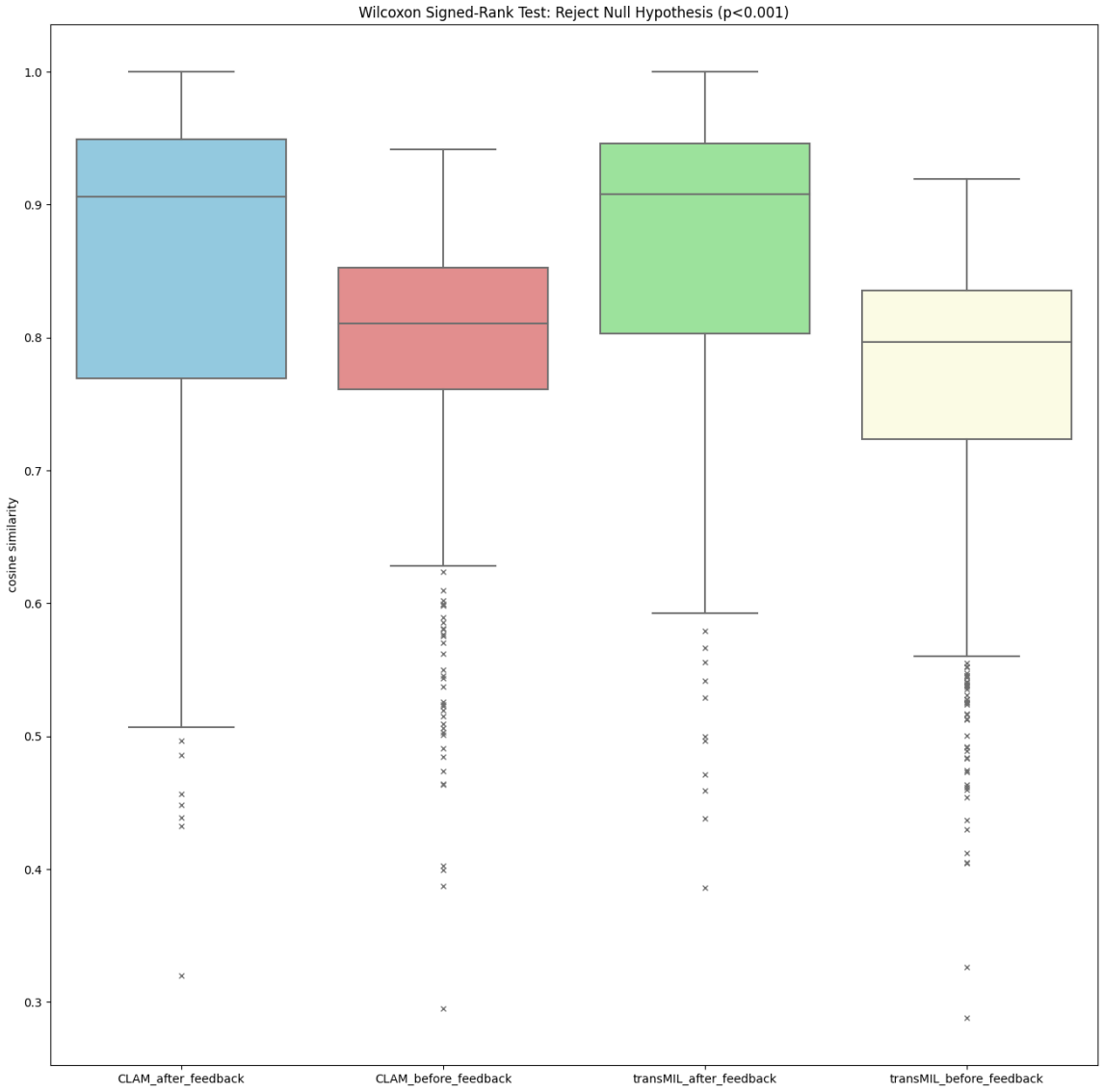



Figure 5.4: Box plot illustrating the distributions of cosine similarities between attention scores generated by CLAM / TransMIL and the expected scores, both before and after expert feedback.

between the cosine similarity before and after the feedback.

It's worth noting that the pathologist's decision to employ the *Flip* operation for most WSIs, around 84.509% (as shown in **Table 5.2**), underscores the counter-intuitive of giving prominence to non-malignant components instead of malignant components. Upon flipping the heatmaps, the pathologist fine-tunes an average area of 16.708% of a flipped WSI by drawing multiple bounding boxes based on the preference to increase, decrease, or maintain the existing scores. The pathologist retains the initial heatmap orientations for the remaining 15.4910% (as shown in **Table 5.3**) of the WSI and engages in a similar annotation process, covering an average area of 14.436% of a WSI.

Table 5.2: Average area finetuned by the pathologist for WSIs that required *Flip* operation

| Criteria | Value |
|---|---|
| Number of WSIs that required the *Flip* operation | 84.509% |
| Average number of patches in such WSIs | 33412 |
| Annotation percentage | 16.708% |
| Average area occupied by *Disagree-2* box on such a WSI | 11.824% |
| Average area occupied by *Disagree-1* box on such a WSI | 0.441% |
| Average area occupied by *Neurtral 0* box on such a WSI | 0.729% |
| Average area occupied by *Disagree+1* box on such a WSI | 2.766% |
| Average area occupied by *Disagree+2* box on such a WSI | 0.946% |

Our findings are coherent with the current literature where our approach demonstrates comparable performance to that reported in [5] that achieves an average AUC of 0.85 on TCGA-BR[2] subtyping. Our RS dataset encompasses approximately 700 images, slightly smaller than TCGA-BR, which includes approximately 1100 images with breast cancer lobular and ductal carcinomas. This similar performance suggests

[2]https://www.cancer.gov/ccg/research/genome-sequencing/tcga/studied-cancers

Table 5.3: Average area annotated by the pathologist for WSIs that did not require *Flip* operation

| Criteria | Value |
|---|---|
| Number of WSIs that did not require the *Flip* operation | 15.4910% |
| Average number of patches in such WSIs | 35446 |
| Annotation percentage | 14.436% |
| Average area occupied by *Disagree-2* box on such a WSI | 9.312% |
| Average area occupied by *Disagree-1* box on such a WSI | 0.370% |
| Average area occupied by *Neurtral 0* box on such a WSI | 0.517% |
| Average area occupied by *Disagree+1* box on such a WSI | 3.279% |
| Average area occupied by *Disagree+2* box on such a WSI | 0.958% |

that our approach is consistent with the state-of-the-art methods described in the literature.

In addition, the marginally diminished performance of CLAM [1] and TransMIL [2] without expert feedback could partly result from using the off-the-shelf Image-Net [25] pre-trained ResNet50 network for feature extraction, not optimally tailored for the histopathological recurrence score tasks.

## 5.4 Heatmaps Analysis on Training samples

In the final phase of our study, we examine the heatmaps generated by the attention-based models before and after the integration of expert feedback. This analysis aims to provide a deeper understanding of the enhancements brought about by our proposed approach.

We focus on four cases, each offering distinct insights into the efficacy of the feedback process on both CLAM [1] and TransMIL [2] performance. These cases serve as pivotal examples that help elucidate the impact of our approach on any attention model's behaviour and heatmap generation. A detailed exploration of specific cases allows us to highlight the significance and relevance of our approach.

### 5.4.1 Case 1- Correcting classification errors

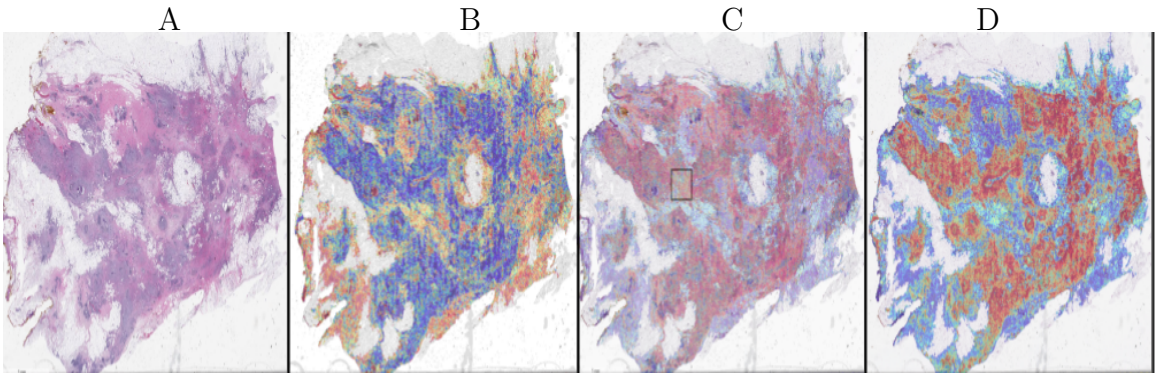



Figure 5.5: **Case 1- CLAM**
**A**: Original WSI; **B**: Heatmap generated by vanilla CLAM; **C**: universal model-independent representation of **A**; **D**: Heatmap generated by CLAM after retrained with the expected scores from **C**
**NOTE:** In **C**, we have *omitted* the procedure of *patch coordinate mapping and thin spline interpolation* and only displayed the GUI annotations as *universal model-independent representation* (as opposed to Fig. 5.1) for the sake of visual understanding and space constraints.

#### CLAM

In this scenario (**Fig. 5.5**), The initial heatmap generated before receiving feedback is contrary to the pathologist's expectations and coincidentally is misclassified. The attention-based model seems to be focused on irrelevant areas while neglecting crucial tumour content (similar to the *negative picture* of attention). Uninformed about the classification outcome, the pathologist, guided by visual evidence, promptly uses our user-friendly Flip button, mirroring the heatmap to access its reversed version.

The expert further refines the heatmap using various bounding box modes. After retraining the CLAM model with the acquired feedback, we observe that the model correctly classifies the WSI, now accurately recognizing essential features. The case initially misclassified as *high* risk with a prediction probability of 0.93, is rectified as *low* risk with a prediction probability of 0.70.

#### TransMIL

Similarly (**Fig. 5.6**), the vanilla model erroneously categorizes the WSI as *high* with a class probability of 0.68. However, subsequent retraining of the model using the

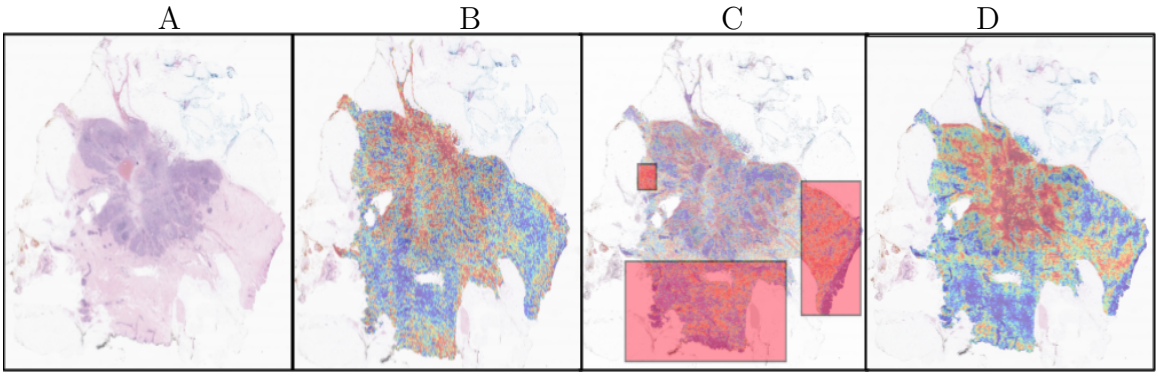



Figure 5.6: **Case 1- TransMIL**
**A**: Original WSI; **B**: Heatmap generated by vanilla TransMIL; **C**: universal model-independent representation of **A**; **D**: Heatmap generated by TransMIL after retrained with the expected scores from **C**
**NOTE:** In **C**, we have *omitted* the process of *patch coordinate mapping and thin spline interpolation* and only displayed the GUI annotations as *universal model-independent representation* (as opposed to Fig. 5.1) for the sake of visual understanding and space constraints.

expected scores derived from the standard reference rectifies this, resulting in the correct classification of the WSI as *low* with an increased probability of 0.80. This case showcases the pivotal role of expert feedback and our proposed approach in rectifying classification errors made by attention-based models.

### 5.4.2 Case 2- Correct classification and finetuning attention map

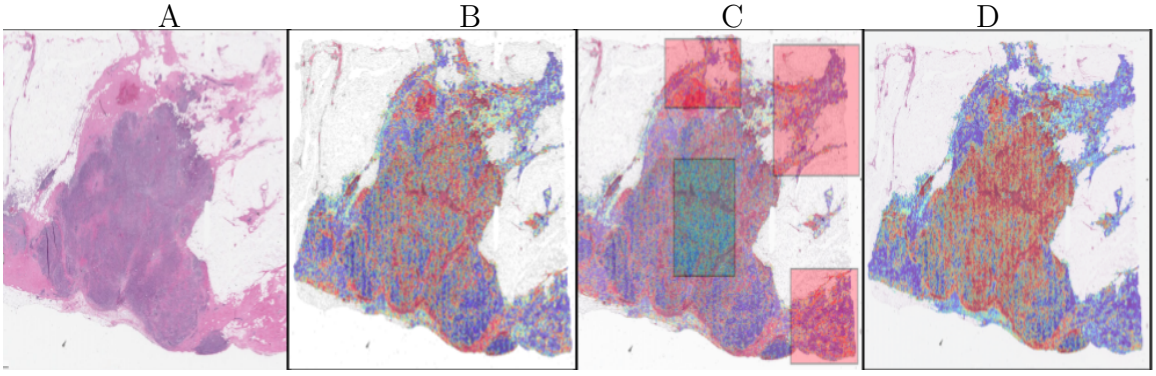



Figure 5.7: **Case 2- CLAM**
**A**: Original WSI; **B**: Heatmap generated by vanilla CLAM; **C**: universal model-independent representation of **A**; **D**: Heatmap generated by CLAM after retrained with the expected scores from **C**
**NOTE:** In **C**, we have *omitted* the process of *patch coordinate mapping and thin spline interpolation* and only displayed the GUI annotations as *universal model-independent representation* (as opposed to Fig. 5.1) for the sake of visual understanding and space constraints.

### CLAM

Here (**Fig. 5.7**), the heatmap generated initially is subpar, yet the model's classification is accurate. It predominately attends to the WSI's borders and outer edges, neglecting significant features.

During the feedback session, the pathologist engages in minor fine-tuning, providing valuable insights for the model. Subsequent retraining redirects the model's attention to the relevant crucial features, resulting in an improved heatmap. Notably, in this *low* class case, the class prediction probability increases from 0.59 to 0.82.

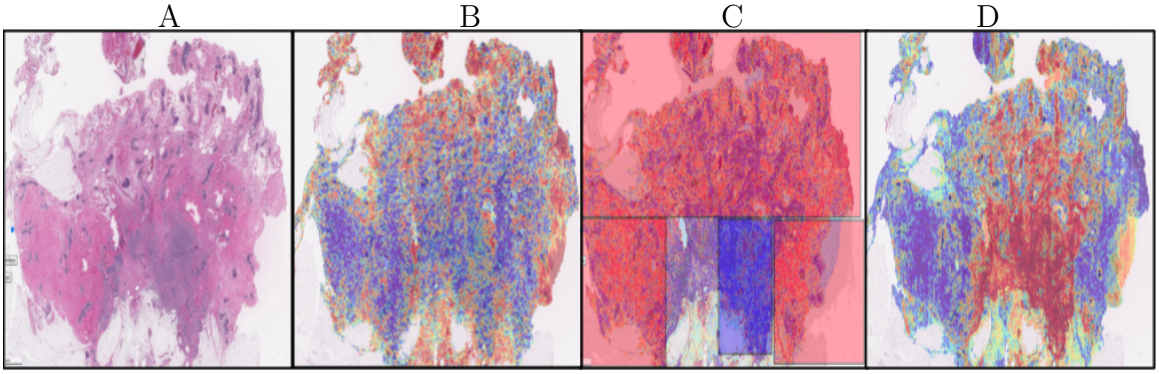



Figure 5.8: **Case 2- TransMIL**
**A**: Original WSI; **B**: Heatmap generated by vanilla TransMIL; **C**: universal model-independent representation of **A**; **D**: Heatmap generated by TransMIL after retrained with the expected scores from **C**
**NOTE:** In **C**, we have *omitted* the process of *patch coordinate mapping and thin spline interpolation* and only displayed the GUI annotations as *universal model-independent representation* (as opposed to Fig. 5.1) for the sake of visual understanding and space constraints.

### TransMIL

The vanilla model (**Fig. 5.8**) correctly classifies the WSI as *low*-risk with a probability of 0.59. However, the subsequent retraining process, fueled by the expected scores derived from the standardized representation, further polishes the model's performance. As a result of this optimization, the prediction probability increases to 0.78 with an improvement in the corresponding heatmap.

These observations emphasize the dual benefits of our feedback-based approach - it corrects classification errors and improves the attention map's alignment with expert

expectations.

### 5.4.3 Case 3- Incorrect classification but improving the correct class prediction probability

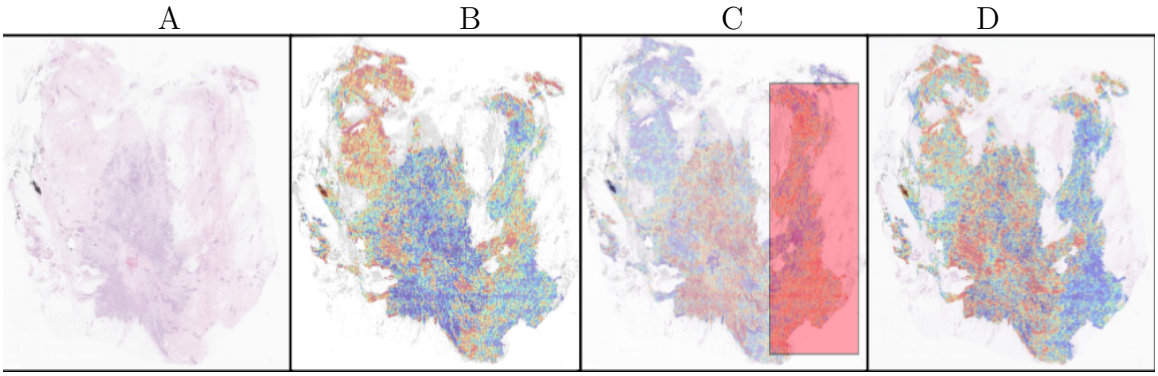



Figure 5.9: **Case 3- CLAM**
**A**: Original WSI; **B**: Heatmap generated by vanilla CLAM; **C**: universal model-independent representation of **A**; **D**: Heatmap generated by CLAM after retrained with the expected scores from **C**
**NOTE:** In **C**, we have *omitted* the process of *patch coordinate mapping and thin spline interpolation* and only displayed the GUI annotations as *universal model-independent representation* (as opposed to Fig. 5.1) for the sake of visual understanding and space constraints.

#### CLAM

In a manner analogous to Case 2, this scenario (**Fig. 5.9**) features a subpar heatmap with an incorrect classification (*high*-risk classified as *low*-risk). Following minor annotations by the pathologist and subsequent model retraining, the heatmap quality improves as it pays less head to artifacts like contours and void spots. Despite this improvement, the case's classification remains erroneous.

However, we notice that the prediction probability of *high* increases from 0.04 to 0.30, indicating that the model now understands some additional representative features of the class by a spike in the class prediction probability.

#### TransMIL

In parallel, the TransMIL model (**Fig. 5.10**) initially misclassifies the WSI as *low*-risk with a class probability of 0.21 for the *high*-risk class. Even after retraining

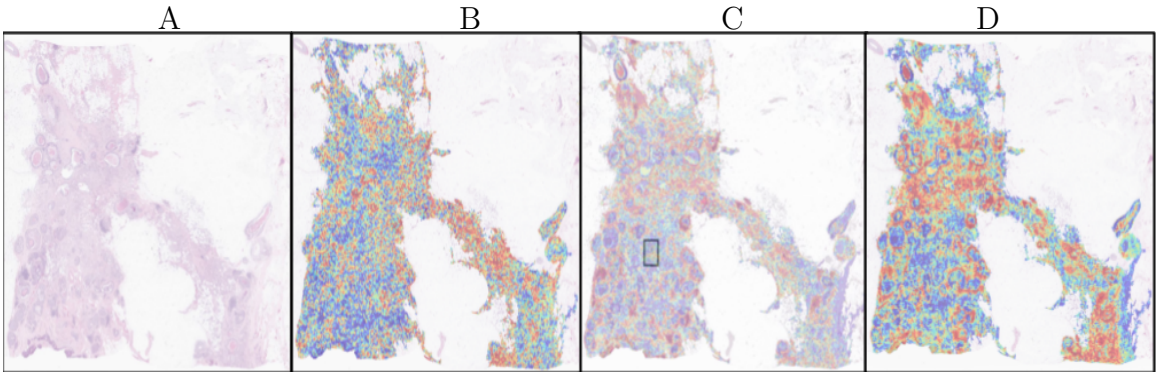



Figure 5.10: **Case 3- TransMIL**
**A**: Original WSI; **B**: Heatmap generated by vanilla TransMIL; **C**: universal model-independent representation of **A**; **D**: Heatmap generated by TransMIL after retrained with the expected scores from **C**
**NOTE:** In **C**, we have *omitted* the process of *patch coordinate mapping and thin spline interpolation* and only displayed the GUI annotations as *universal model-independent representation* (as opposed to Fig. 5.1) for the sake of visual understanding and space constraints.

the model using the attention scores from the GUI feedback, the WSI's classification persists as *low*, albeit with an increased prediction probability for the *high* class, now standing at 0.40.

Despite the classification not changing, the increased prediction probability indicates the model's improved capacity to recognize *high*-risk features, signifying the positive impact of our feedback-driven approach.

### 5.4.4 Case 4- Correct classification by improving correct class prediction probability for borderline cases

We regard WSIs with classification outcomes of prediction probabilities ranging from 0.45 to 0.55 as cases situated in the borderline region.

#### CLAM

Within this context (**Fig. 5.11**), we notice that the initial and the post-feedback heatmaps are similar. It means that there are some refinements through fine-tuning by our pathologist. While analyzing the classification results, we notice that the case was misclassified as *low* initially.

The post-feedback heatmap displays slight improvements due to a minor elevation

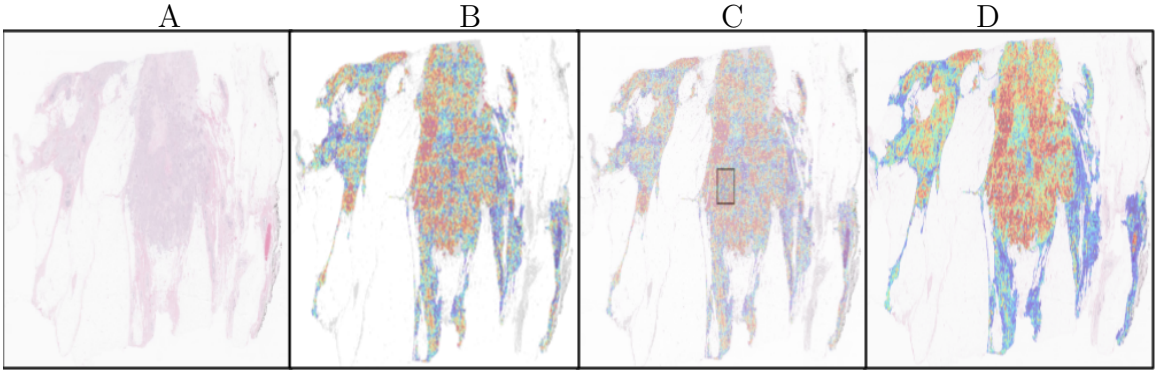



Figure 5.11: **Case 4- CLAM**
**A**: Original WSI; **B**: Heatmap generated by vanilla CLAM; **C**: universal model-independent representation of **A**; **D**: Heatmap generated by CLAM after retrained with the expected scores from **C**
**NOTE:** In **C**, we have *omitted* the process of *patch coordinate mapping and thin spline interpolation* and only displayed the GUI annotations as *universal model-independent representation* (as opposed to Fig. 5.1) for the sake of visual understanding and space constraints.

in the class probability from 0.40 to 0.54. This improvement leads to the accurate reclassification of the case as *high*-risk.

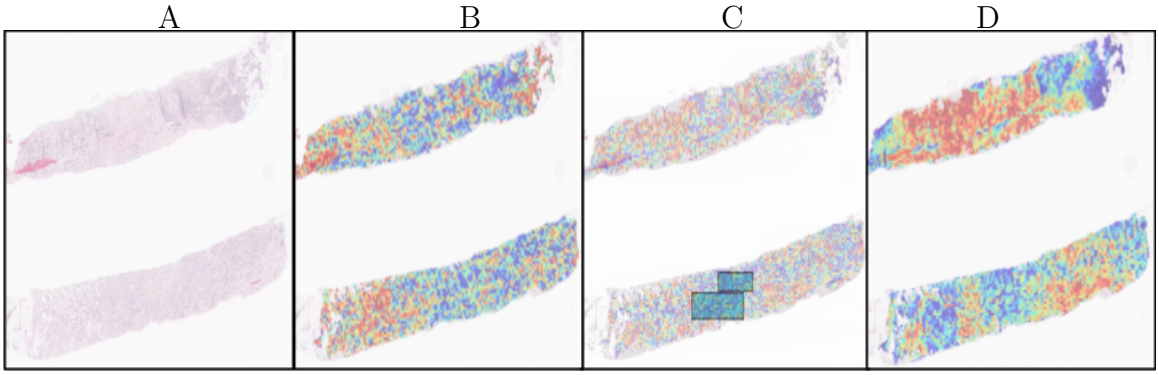



Figure 5.12: **Case 4- TransMIL**
**A**: Original WSI; **B**: Heatmap generated by vanilla TransMIL; **C**: universal model-independent representation of **A**; **D**: Heatmap generated by TransMIL after retrained with the expected scores from **C**
**NOTE:** In **C**, we have *omitted* the process of *patch coordinate mapping and thin spline interpolation* and only displayed the GUI annotations as *universal model-independent representation* (as opposed to Fig. 5.1) for the sake of visual understanding and space constraints.

### TransMIL

The initial classification (**Fig. 5.12**) of the WSI as *high* with a prediction probability of 0.54 changes after re-training the model using the standard reference. Subsequently, the case is correctly classified as *low* with a prediction probability of 0.81.

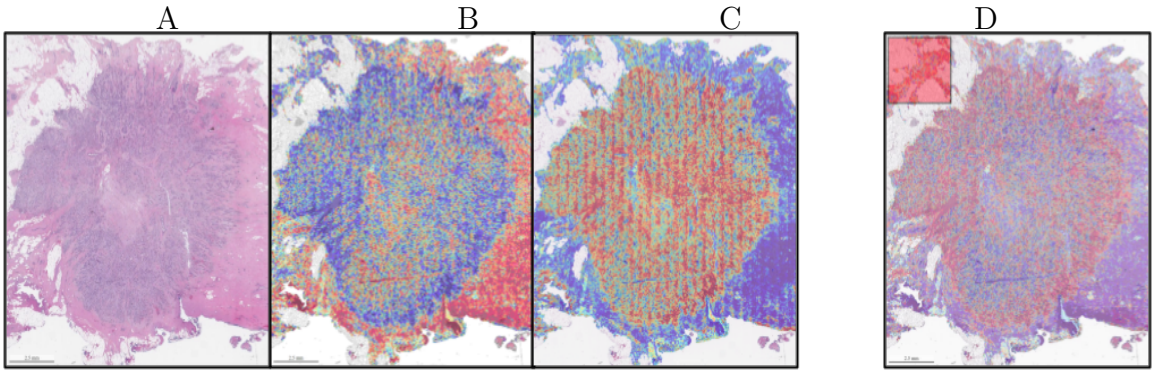



Figure 5.13: **Example I** - Before feedback - classified as *high* (wrong) with its prob. of 0.63
After feedback - classified as *low* (correct) with its prob. of 0.80
**A**: Test WSI; **B**: Heatmap generated by vanilla CLAM; **C**: Heatmap generated by CLAM after trained with expert feedback; **D**: universal model-independent representation of **A** which was not shown to CLAM during testing
**NOTE:** In **D**, we have *omitted* the process of *patch coordinate mapping and thin spline interpolation* and only displayed the GUI annotations as *universal model-independent representation* (as opposed to Fig. 5.1) for the sake of visual understanding and space constraints.

This shift in classification underscores the efficacy of our feedback-driven approach, particularly in cases where initial probabilities lie within the borderline region.

### 5.4.5 Heatmap Analysis on Test samples

Despite retraining the models with feedback, it is evident that some regions of relevant significance continue to be delicately overlooked by the attention-based models (as depicted in **Example I - Fig. 5.13**, **Example II - Fig. 5.14**). This observation is anticipated due to the diverse appearances of artifacts in areas with and without cancer, making their discrimination challenging.

Nevertheless, our method showcases promising potential in enhancing the efficacy of learning models with just a single round of feedback in the form of bounding boxes. The improved performance for most of the cases further suggests the significance of the additional MSE loss term. This term, which penalizes discrepancies between expected and current scores, aids the model in focusing on pertinent information, thereby leading to a better grasp of features and overall predictive performance.

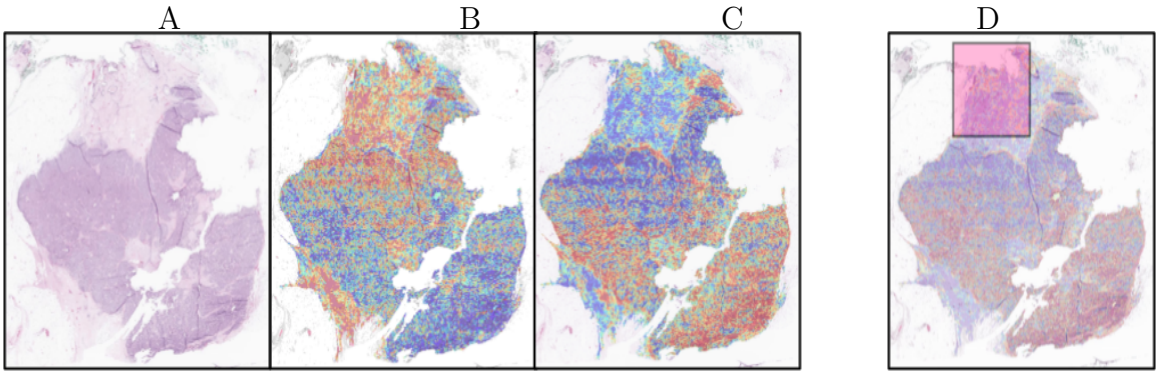



Figure 5.14: **Example II** - Before feedback - classified as *low* (wrong) with *high* prob. of 0.25
After feedback - classified as *low* (still wrong) with an increase in *high* prob. of 0.43
**A**: Test WSI; **B**: Heatmap generated by vanilla CLAM; **C**: Heatmap generated by CLAM after trained with expert feedback; **D**: universal model-independent representation of **A** which was not shown to CLAM during testing
**NOTE:** In **D**, we have *omitted* the process of *patch coordinate mapping and thin spline interpolation* and only displayed the GUI annotations as *universal model-independent representation* (as opposed to Fig. 5.1) for the sake of visual understanding and space constraints.

# Chapter 6

# Conclusion

## 6.1 Recapitulation of Findings

Our focus lies in addressing the distinction between Luminal A and Luminal B BC types, presenting a formidable challenge to AI in predicting OncotypeDX® results based on BC morphology. WSI analysis confronts the *WHERE problem*, aiming to identify regions of significant diagnostic value within slides characterized by an admixture of malignant and non-malignant components randomly dispersed. While the weakly supervised attention mechanism alleviates the *WHERE problem* for metastasis datasets, we observe that for predicting the Recurrence Score of a dataset, the attention mechanism manifests a *negative picture of attention*. In essence, the existing attention has not proven entirely sufficient. To overcome these hurdles, we sought the guidance of a clinical expert to rectify the initial attention heatmaps generated by CLAM and TransMIL. Even after training the model with expert feedback, we still note instances where tumour components evade the model's attention due to the diverse appearances of artifacts in areas with and without cancer. It's worth noting that we've performed this feedback process only for a single iteration.

On a positive note, we've observed significant improvements in heatmap quality and quantitative results. We note an enhancement in Recurrence Score (RS) prediction performance after retraining both models with input from the pathologist's annotations. Specifically, there is a significant 5% increase in the validation-test Area Under

the Curve (AUC) and a 4% rise in validation-test accuracy for the CLAM model. Similarly, for the TransMIL model, the validation-test AUC improves by 4.5%, accompanied by a 3% boost in validation-test accuracy. We also observe an increase in cosine similarity between the pathologist's GUI-based attention scores and the attention maps produced by the trained models. Specifically, we notice a 5% increase for the CLAM model and an impressive 10% rise for the TransMIL model, clearly indicating the effectiveness of expert feedback in enhancing the alignment between human annotations and machine-generated attention patterns.

## 6.2 Unexplored Avenues

In a model's learning process, regularization techniques like L1 or L2 usually prevent overfitting and improve performance and generalization. Our finding indicates that the improvement in the performance of CLAM [1] and TransMIL [2] from adding the extra MSE loss term cannot be attributed solely to such conventional regularization strategies. Instead, we suspect that this performance improvement arises from the substantial contributions of the attention sub-module. Latent representations refer to hidden and abstract features a model learns to extract from the data. These representations are often critical for understanding complex patterns and making accurate predictions. It would be worthwhile to experiment if this sub-module significantly facilitates more effective latent space representation learning in our experiments, consequently fostering performance enhancement. This exploration could provide valuable insights into the mechanics of attention-based models like CLAM and TransMIL and help us better understand the mechanisms driving their success.

Our focus primarily revolves around investigating the integration of expert feedback into attention-based models to achieve enhanced performance on our in-house dataset. While our study has adopted a human-supervised approach, we acknowledge the potential merits of strategies such as transfer learning from publicly available datasets and unsupervised domain adaptation. Although we do not delve into their

potential benefits in our study, these strategies hold promise in refining and expanding the capabilities of our dynamic and adaptive approach with a pathologist's supervision. These strategies have the potential to offer fresh perspectives and meaningful contributions to the evolving field.

## 6.3 Takeaways

Our research findings strongly accentuate the impact of expert feedback integration through an intuitive GUI on the performance of attention-based models. The discernible boosts in validation-test AUC and accuracy metrics provide compelling evidence of the human-in-the-loop interaction in refining AI models purpose-built for medical image analysis.

This study serves the potent synergy through the collaboration between AI and domain experts by harmonizing learning algorithms with the nuanced insights of clinicians. This collaborative paradigm holds immense promise for surmounting the challenges inherent in medical image analysis, propelling the field toward new horizons of research and innovation with better prediction metrics values that can be incorporated into the clinical workflows.

# Bibliography


[1] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature biomedical engineering*, vol. 5, no. 6, pp. 555–570, 2021.

[2] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, *et al.*, "Transmil: Transformer based correlated multiple instance learning for whole slide image classification," *Advances in neural information processing systems*, vol. 34, pp. 2136–2147, 2021.

[3] J. A. Sparano *et al.*, "Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer," *New England Journal of Medicine*, vol. 379, no. 2, pp. 111–121, 2018.

[4] Y. Liang, M. Li, and C. Jiang, "Generating self-attention activation maps for visual interpretations of convolutional neural networks," *Neurocomputing*, vol. 490, pp. 206–216, 2022.

[5] Z. Gao *et al.*, "A semi-supervised multi-task learning framework for cancer classification with weak annotation in whole-slide images," *Medical Image Analysis*, vol. 83, p. 102 652, 2023.

[6] S. Paik *et al.*, "A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer," *New England Journal of Medicine*, vol. 351, no. 27, pp. 2817–2826, 2004.

[7] E Senkus *et al.*, "Primary breast cancer: Esmo clinical practice guidelines for diagnosis, treatment and follow-up," *Annals of oncology*, vol. 26, pp. v8–v30, 2015.

[8] J. Webster and R. Dunstan, "Whole-slide imaging and automated image analysis: Considerations and opportunities in the practice of pathology," *Veterinary pathology*, vol. 51, no. 1, pp. 211–223, 2014.

[9] U. Catalyurek, M. D. 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.

[10] R. Ferreira *et al.*, "The virtual microscope.," in *Proceedings of the AMIA Annual Fall Symposium*, American Medical Informatics Association, 1997, p. 449.

[11] C. Daniel *et al.*, "Recent advances in standards for collaborative digital anatomic pathology," *Diagnostic pathology*, vol. 6, pp. 1–10, 2011.

[12] J. M. Chang *et al.*, "Back to basics: Traditional nottingham grade mitotic counts alone are significant in predicting survival in invasive breast carcinoma," *Annals of surgical oncology*, vol. 22, pp. 509–515, 2015.

[13] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[14] V. Ojansivu and J. Heikkilä, "Blur insensitive texture classification using local phase quantization," in *Image and Signal Processing: 3rd International Conference, ICISP 2008. Cherbourg-Octeville, France, July 1-3, 2008. Proceedings 3*, Springer, 2008, pp. 236–243.

[15] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification.," *Journal of machine learning research*, vol. 10, no. 2, 2009.

[16] A. Tharwat, "Linear vs. quadratic discriminant analysis classifier: A tutorial," *International Journal of Applied Pattern Recognition*, vol. 3, no. 2, pp. 145–180, 2016.

[17] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pp. 144–152.

[18] V. Lepetit and P. Fua, "Keypoint recognition using randomized trees," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 9, pp. 1465–1479, 2006.

[19] A. Chan and J. A. Tuszynski, "Automatic prediction of tumour malignancy in breast cancer with fractal dimension," *Royal Society open science*, vol. 3, no. 12, p. 160 558, 2016.

[20] N. A. Barsha, A. Rahman, and M. Mahdy, "Automated detection and grading of invasive ductal carcinoma breast cancer using ensemble of deep learning models," *Computers in Biology and Medicine*, vol. 139, p. 104 931, 2021.

[21] M. A. Jawad and F. Khursheed, "Histo-fusion: A novel domain specific learning to identify invasive ductal carcinoma (idc) from histopathological images," *Multimedia Tools and Applications*, pp. 1–22, 2023.

[22] P. Bankhead *et al.*, "Qupath: Open source software for digital pathology image analysis," *Scientific reports*, vol. 7, no. 1, pp. 1–7, 2017.

[23] A. E. Carpenter *et al.*, "Cellprofiler: Image analysis software for identifying and quantifying cell phenotypes," *Genome biology*, vol. 7, pp. 1–11, 2006.

[24] K. Cizkova, T. Foltynkova, M. Gachechiladze, and Z. Tauber, "Comparative analysis of immunohistochemical staining intensity determined by light microscopy, imagej and qupath in placental hofbauer cells," *Acta histochemica et cytochemica*, vol. 54, no. 1, pp. 21–29, 2021.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[26] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *International conference on machine learning*, PMLR, 2018, pp. 2127–2136.

[27] B Krishnakumar and K Kousalya, "Deep learning techniques for breast cancer diagnosis: A systematic review," in *International Conference on Innovative Technology, Engineering and Science*, Springer, 2020, pp. 155–171.

[28] G. Campanella *et al.*, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature medicine*, vol. 25, no. 8, pp. 1301–1309, 2019.

[29] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[30] M. Zhuang, Z. Chen, Y. Yang, L. Kettunen, and H. Wang, "Annotation-efficient training of medical image segmentation network based on scribble guidance in difficult areas," *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–10, 2023.

[31] T. T. Brunyé, E. Mercan, D. L. Weaver, and J. G. Elmore, "Accuracy is in the eyes of the pathologist: The visual interpretive process and diagnostic accuracy with digital whole slide images," *Journal of biomedical informatics*, vol. 66, pp. 171–179, 2017.

[32] E. Mercan, L. G. Shapiro, T. T. Brunyé, D. L. Weaver, and J. G. Elmore, "Characterizing diagnostic search patterns in digital breast pathology: Scanners and drillers," *Journal of digital imaging*, vol. 31, pp. 32–41, 2018.

[33] H. Le, D. Samaras, T. Kurc, R. Gupta, K. Shroyer, and J. Saltz, "Pancreatic cancer detection in whole slide images using noisy label annotations," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22*, Springer, 2019, pp. 541–549.

[34] Y. B. Can, K. Chaitanya, B. Mustafa, L. M. Koch, E. Konukoglu, and C. F. Baumgartner, "Learning to segment medical images with scribble-supervision alone," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, Springer, 2018, pp. 236–244.

[35] N. G. Laleh *et al.*, "Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology," *Medical image analysis*, vol. 79, p. 102 474, 2022.

[36] T. F. Sterkenburg and P. D. Grünwald, "The no-free-lunch theorems of supervised learning," *Synthese*, vol. 199, no. 3-4, pp. 9979–10 015, 2021.

[37] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[38] *Choosing colormaps in matplotlib - matplotlib 3.7.2 documentation*. [Online]. Available: https://matplotlib.org/stable/tutorials/colors/colormaps.html.

[39] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial intelligence*, vol. 89, no. 1-2, pp. 31–71, 1997.

[40] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," *Pattern Recognition*, vol. 74, pp. 15–24, 2018.

[41] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[42] Openseadragon, *Openseadragon/openseadragon: An open-source, web-based viewer for zoomable images, implemented in pure javascript*. [Online]. Available: https://github.com/openseadragon/openseadragon.

[43] A. Goode, B. Gilbert, J. Harkes, D. Jukic, and M. Satyanarayanan, "Openslide: A vendor-neutral software foundation for digital pathology," *Journal of pathology informatics*, vol. 4, no. 1, p. 27, 2013.

[44] *Welcome to flask - flask documentation (2.3.x)*. [Online]. Available: https://flask.palletsprojects.com/en/2.3.x/.

[45] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2623–2631.

[46] *American cancer society*. [Online]. Available: https://www.cancer.org/research/cancer-facts-statistics/breast-cancer-facts-figures.html.

[47] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer statistics, 2023," *CA: a cancer journal for clinicians*, vol. 73, no. 1, pp. 17–48, 2023.

[48] B. Turner, M. Sanders, A Breaux, A Soukiazian, N Soukiazian, and D. Hicks, "Abstract p2-08-24: The average modified magee score can be helpful in predicting an oncotype dx recurrence score 25," *Cancer Research*, vol. 79, no. 4_Supplement, P2–08, 2019.

[49] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, and Á. Fernández-Leal, "Human-in-the-loop machine learning: A state of the art," *Artificial Intelligence Review*, vol. 56, no. 4, pp. 3005–3054, 2023.

[50] Y. Xiong *et al.*, "Nyströmformer: A nyström-based algorithm for approximating self-attention," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 14 138–14 148.

[51] X. Wang *et al.*, "Weakly supervised deep learning for whole slide lung cancer image analysis," *IEEE transactions on cybernetics*, vol. 50, no. 9, pp. 3950–3962, 2019.

[52] N. Dimitriou, O. Arandjelović, and P. D. Caie, "Deep learning for whole slide image analysis: An overview," *Frontiers in medicine*, vol. 6, p. 264, 2019.