UNIVERSITY OF TURKU

Post-hoc Interpretation of Deep Learning-based Survival Model on Colorectal Cancer Using Hematoxylin and Eosin Slides

Master’s Degree Programme in Biomedical Imaging

Master's thesis

Le Thi Thanh Phuong

28.04.2026

Turku

Master's thesis

Subject: Master's Degree Programme in Biomedical Imaging

Author(s): Le Thi Thanh Phuong

Title: Post-hoc Interpretation of Deep Learning-based Survival Model on Colorectal Cancer Using Hematoxylin and Eosin Slides

Supervisor(s): Professor Pekka Ruusuvuori; Docent Camilla Böckelman

Number of pages: 78 pages

Date: 28.04.2026

Abstract

Background: Colorectal cancer (CRC) prognosis relies primarily on the Tumor-Nodes-Metastasis (TNM) staging system, yet tumor heterogeneity limits its accuracy. Therefore, numerous new pathological features and biomarkers have been developed to complement this system. Nevertheless, they have been overcomplicated to implement clinically. Deep learning (DL) models show promising ability to solve this issue by extracting prognostic information from routine hematoxylin and eosin (H&E)-stained whole slide images (WSIs). However, their clinical adoption remains limited because most models lack explanation, while clinical pathologists require interpretability.

Objective: The purpose of this thesis was to interpret a multiple-instance learning (MIL)-based survival model for CRC, addressing three research questions: (1) model performance, (2) histopathological patterns within the image patches most strongly influencing the model's predictions, and (3) sources of model failure.

Methods: A Clustering-constrained Attention Multiple Instance Learning (CLAM) model using UNI2-h foundation model embeddings predicted 5-year survival status, approximating disease-specific survival (DSS), as non-disease-related deaths within 5 years were excluded, from WSIs in a single-center, retrospective cohort. Post-hoc interpretation combined attention heatmaps, K-means clustering (k=75) of high-confidence tile embeddings, forward stepwise selection of prognostic clusters, and pathologist review of representative tiles from the top 10 most informative feature clusters.

Results: The UNI2-h–CLAM survival model achieved robust discrimination of 5-year survival status in CRC, with AUC = 0.79 ± 0.034 at the patient-level, c-index = 0.738. The Kaplan-Meier analysis proved a clear survival stratification by CLAM predictions, with a hazard ratio (HR) of 4.27 (95% CI, 3.22–5.66; log-rank p-value <0.001). The model mostly focused on the tumor epithelium in alive-predicted cases, whereas invasion zones and peritumoral stroma were important in dead-predicted cases. Dead-

predicted tiles showed high-grade atypia, poorly differentiated clusters, immature/myxoid desmoplastic stroma, and stroma-high feature, while alive-predicted tiles displayed preserved glands and inflammatory infiltrates. The top 10 clusters explained 73.9% of model variance (R^2=0.739). Errors involved non-representative WSIs selection, intra-patient heterogeneity, artifacts, and tissue microarray core loss.

Conclusion: This multiple-instance learning-based DL model predicts CRC survival effectively. The model independently learned prognostic features aligning with WHO/AJCC-recognized markers (poor differentiation, tumor budding-like clusters, immature desmoplasia). This data-driven interpretation framework bridges the gap between deep learning models and clinical pathology, supporting tumor-stroma ratio, desmoplastic reaction as robust artificial intelligence - discoverable biomarkers for risk stratification.

Keywords: Survival Prediction, Deep learning, Colorectal Cancer, Digital Pathology.

Table of contents

List of tables

List of figures

Abbreviations

AI	Artificial Intelligence
AIC	Akaike Information Criterion
AJCC	American Joint Committee on Cancer
AUC	Area under the curve
AUC ROC	Area Under the Receiver Operating Characteristic Curve
BRAF	B-Raf Proto-Oncogene, Serine/Threonine Kinase
C-index	Concordance index
CAP	College of American Pathologists
CDX2	Caudal Type Homeobox 2
CI	Confidence interval
CLAM	Clustering-Constrained Attention Multiple Instance Learning
CMS	Consensus Molecular Subtypes
CNN	Convolutional Neural Network
COAD	colon adenocarcinoma
CRC	Colorectal Cancer
CRCRS	Colorectal Cancer Risk Score
CTCs	Circulating Tumor Cells
ctDNA	Circulating Tumor Deoxyribonucleic Acid
DACHS	Darmkrebs: Chancen der Verhütung durch Screening
DeepDisMISL	Distribution based multiple-instance survival learning algorithm
DL	Deep Learning
DLS	Deep Learning System
DR	Desmoplastic Reaction

DSS	Disease – Specific Survival
DUSSEL	Düsseldorf cohort
EGFR	Epidermal Growth Factor Receptor
FIMEA	Finnish Medicines Agency
Grad-CAM	Gradient – Weighted Class Activation Mapping)
H&E	Hematoxylin and Eosin
HER2	Human Epidermal Growth Factor Receptor 2
HR	hazard ratio
HUS	Helsinki University Hospital (Helsingin yliopistollinen sairaala)
KRAS	Kirsten Rat Sarcoma Viral Oncogene Homolog
LSTM	Long Short – Termmemory
LVI	Lymphovascular Invasion
MCO	Molecular and Cellular Oncology cohort
MIL	Multiple Instance Learning
MMR	Mismatch Repair
MSI	Microsatellite Instability
OS	Overall Survival
PLCO	Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial
PNI	Perineural Invasion
PRISM	Prognostic Representation of Integrated Spatial Morphology
px	pixels
READ	rectal adenocarcinoma
RFS	relapse-free survival
SD	Standard deviation
SHAP	Shapley Additive Explanations

SSL	Self-Supervised Learning
SVD	singular value decomposition
t-SNE	t-distributed Stochastic Neighbor Embedding
TCGA	The Cancer Genome Atlas
TIL	Tumor-Infiltrating Lymphocyte
TMA	Tissue Microarray
TMB	Tumor Mutation Burden
TME	Tumor Microenvironment
TNM	Tumor – Node – Metastasis
TSR	Tumor – Stroma Ratio
UICC	Union for International Cancer Control
WHO	World Health Organization
WSI	Whole Slide Image
WSU	Wayne State University cohort

1 INTRODUCTION

Globally, colorectal cancer (CRC) ranks as the third most frequently diagnosed malignancy, accounting for nearly one-tenth of total cancer incidence. It is also the second most common cause of cancer-related mortality worldwide [1, 2]. Therefore, besides the precise diagnosis requirement, colorectal cancer prognosis plays a crucial role in choosing treatment therapies and follow-up plans. Particularly, in the personalized medicine era, individual prognosis leads to distinct treatment strategies [3, 4].

In clinical practice, therapeutic decision-making in CRC is primarily guided by the tumor-node-metastasis (TNM) staging system, supplemented with additional prognostic factors [5]. However, achieving accurate prognostication remains a challenge due to tumor heterogeneity and complex clinical presentations, and the TNM system fails to include all the diversity in tissue slides [5, 6].

Research in the past few decades has sought additional markers that can better predict clinical outcomes and help tailor treatment strategies for patients [4, 5]. Therefore, there is a need to integrate non-anatomical factors, such as tumor budding, extra nodal tumor deposits, molecular markers, and treatment-related changes, into staging systems. However, this might be overcomplicated and limit the practical implementation [7].

In that scenario, deep learning techniques based on hematoxylin and eosin (H&E)-stained images have gained significant attention as a promising tool for extracting prognostic information. Although artificial intelligence (AI) has shown potential in pathology, it is still not widely accepted because of its lack of interpretability. Pathologists need not only accurate but also explainable models that reveal which tissue regions drive predictions [8]. In some previous studies, authors tried different methods, such as using Gradient-weighted class activation mapping (Grad-CAM), heatmaps, or clustering, to interpret the model [4, 9, 10]. Nonetheless, over the past years, several studies have accepted latent features without interpretation [3, 11], highlighting a lack of interpretability in existing approaches. This proves that interpreting machine-derived prognostic features remains a big challenge [8, 12].

To address these limitations, multiple instance learning (MIL) frameworks have emerged as a powerful weakly supervised paradigm for whole slide image (WSI) analysis, requiring only

slide-level labels and bypassing the need for costly patch-level annotations that traditional Convolutional Neural Networks (CNNs) demand [13]. Among MIL approaches, Clustering-constrained Attention Multiple Instance Learning (CLAM) has demonstrated strong data efficiency, adaptability across cancer types, and critically built-in interpretability through attention-based identification of diagnostically relevant sub-regions within WSIs [13]. Concurrently, pathology foundation models pre-trained via Self-Supervised Learning (SSL) on large-scale histopathology datasets have demonstrated superior feature representations compared to conventional ImageNet-pretrained CNNs, capturing domain-specific morphological patterns that substantially improve downstream task performance [14, 15]. Building on these developments, in this thesis, a two-stage strategy was adopted: using the UNI2-h foundation model for patch-level feature extraction, followed by CLAM for attention-based survival prediction. Beyond spatial interpretability provided by CLAM's attention heatmaps, the top-attended tiles were clustered using the K-means algorithm to uncover recurring morphological phenotypes associated with mortality, bridging the gap between model-identified prognostic regions and human-interpretable tissue patterns [16].

2 OBJECTIVES AND RESEARCH QUESTIONS

This thesis aims first to evaluate an MIL-based survival model for CRC. Secondly, it investigates the outputs to identify the most influential features driving outcome predictions. The goal is to provide histologically relevant explanations to support pathologists' understanding, possibly explore new features that the human eye skips, and be more confident in implementing them clinically. To achieve that general goal, three specific research questions were posed:

1. How do the deep learning (DL) prediction model's outputs compare with the ground-truth 5-year outcomes?

2. Which patterns/ structures most strongly drive survival model predictions?

3. What regions/patterns does the DL model fail on, and do these align with the challenge cases also for the pathologists?

3 Background

3.1 Colorectal cancer histopathology

3.1.1 Epidemiology and the importance of early prognosis

Epidemiology of colorectal cancer

Colorectal cancer (CRC) is a major global health challenge, ranking as the third most commonly diagnosed malignancy and the second leading cause of cancer death worldwide [1, 2]. In 2022, there were more than 1.9 million new CRC cases and 904, 019 deaths occurred. The global burden of CRC is expected to rise significantly, with projected incidence and mortality reaching 2.36 million and 1.11 million cases, respectively, by 2050 across both sexes. Males had higher incidence and mortality rate then females in almost all regions [17]. Among global regions, Asia and Europe are leading in both incidence and mortality, contributing over 50% and 27%, respectively [2]. Eastern Asia accounted for the highest number of both new cases and deaths worldwide [17].

Most CRCs are sporadic and result from multifactorial interactions among genetic predisposition, environmental, and lifestyle factors [6]. Lifestyle risk-factors include obesity, physical inactivity, diets high in red and processed meat and low in fibre, alcohol consumption, and smoking, all of which are becoming more prevalent in developing economies. As a result, the CRC burden is gradually shifting from historically high-incidence Western countries towards low- and middle-income countries undergoing rapid economic and lifestyle change. An additional concern is the rising incidence of CRC before the age of 50 in many high-income countries [6, 18]. Genetic and molecular heterogeneity further complicate the epidemiological picture. Germline alterations such as MLH1 and APC mutations underlie hereditary syndromes (Lynch syndrome, familial adenomatous polyposis), but the majority of cases arise from stepwise accumulation of somatic changes along the adenoma–carcinoma sequence, serrated, or inflammatory pathways. Right-sided, left-sided, and rectal cancers differ in mutation spectra, tumor microenvironment, and microbial profiles, which translate into distinct prognoses and therapeutic responses and motivate location-specific risk stratification strategies [6].

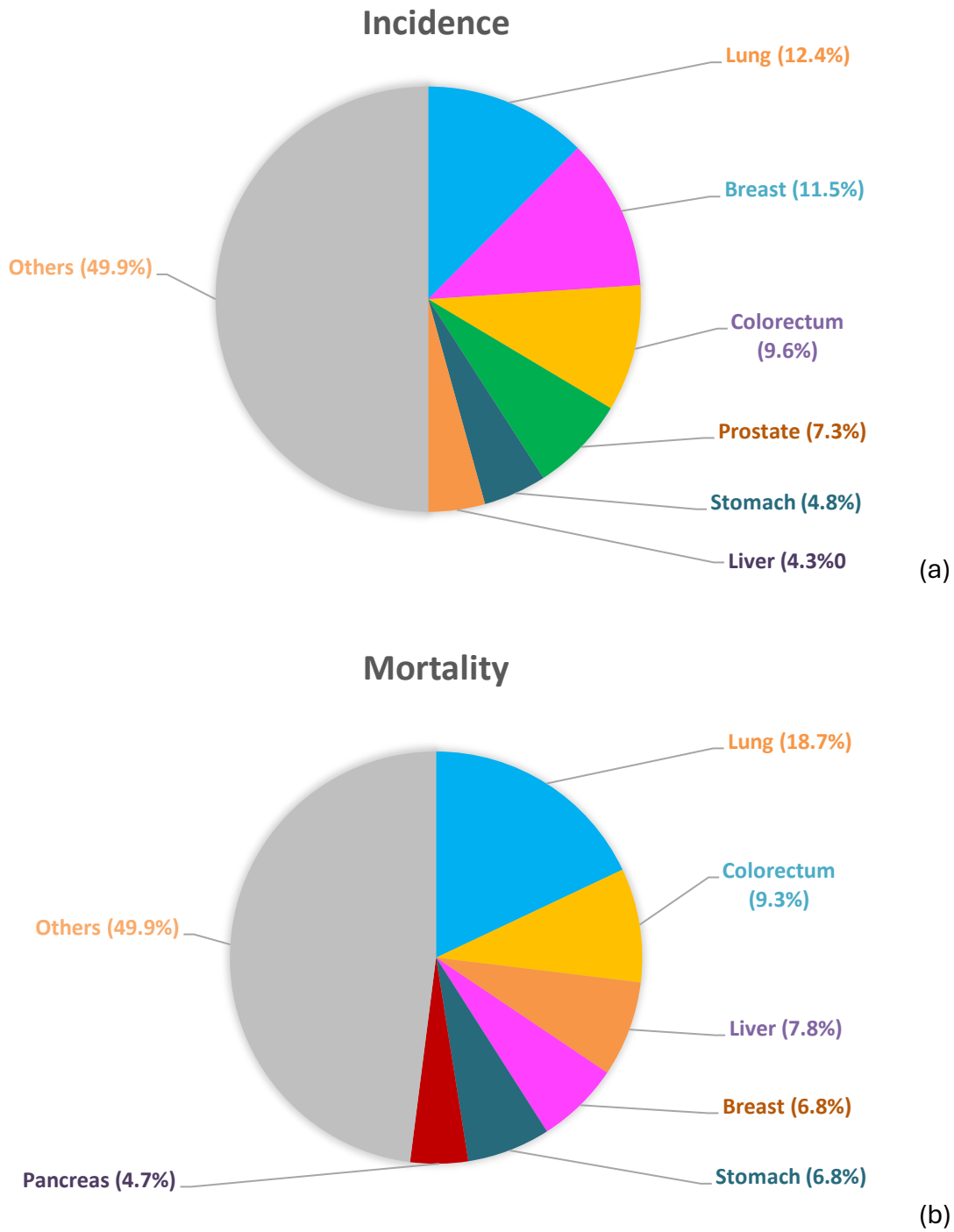


Figure 3.1: Global distribution of cancer incidence (a) and mortality (b) by type, 2022 (modified from Global Cancer Observatory, International Agency for Research on Cancer) [2]

Role of survival prediction

Prognosis plays an essential role across different stages in CRC management [5]. At diagnosis, robust prognostic models help identify patients at high risk of recurrence or death who may benefit from intensified adjuvant chemotherapy or combined-modality treatment,

while sparing low-risk patients from unnecessary toxicity. At a population level, accurate survival estimates underpin health-economic evaluations, planning of oncology services, and evaluation of screening programmes [6].

Clinically, CRC prognosis highly depends on the TNM staging [6, 19]. Five-year relative survival exceeds 90% for early-stage disease but falls to around 15% once distant metastases are present, illustrating the critical importance of early detection and accurate risk assessment [6]. However, TNM staging system alone cannot fully capture all prognostic factors and fails to reflect the clinical heterogeneity of colorectal cancer [19]. Within the same TNM stage, outcomes vary widely owing to differences in tumor biology, treatment tolerance, comorbidities, and host factors, creating a need for more refined prognostic tools that go beyond anatomic staging alone [4, 6, 19].

Taken together, the growing and shifting global burden of colorectal cancer, its marked biological heterogeneity, and the limitations of prognostication systems based solely on anatomical staging make accurate survival prediction a cornerstone of modern CRC care. Accurate risk prediction strategies enable personalised treatment selection, more rational clinical trial design, and informed health-care planning in anticipation of the projected rise in incidence and mortality [6].

3.1.2 Pathological features of colorectal cancer prognosis

Traditional histopathological prognosticators

Pathological prognostication in CRC relies on anatomic stage, designated by the Union for International Cancer Control (UICC) and American Joint Committee on Cancer (AJCC) TNM system, because depth of invasion (T - Tumor), nodal status (N - Nodes), and distant metastasis (M) stratify outcome and guide adjuvant therapy decisions (Table 3.1). However, the TNM staging system cannot capture adequate prognostic features [4, 5, 20]. Significant divergent outcomes persist within the same pathological stage, especially in stage II–III disease, motivating the use of additional risk factors for refined risk stratification [19-21].

Therefore, other microscopic features that have been considered for use in prognosis and CRC management include tumor differentiation (grade), location of tumor (whether in right-sided, left-sided colon or rectum), lymphovascular invasion (LVI), tumor-infiltrating

lymphocytes (TILs), perineural invasion (PNI), and tumor budding at the invasive front [19, 22]. Particularly, tumor budding, which is defined as small clusters of 1 to 4 tumor cells at the invasive front and standardized into BD1–BD3 categories, is an independent adverse marker, with particular value in identifying high-risk stage II disease [19, 22, 23].

Table 3.1: TNM staging for colorectal cancer [24]

Tumor	
Tis	The tumor is limited only in the mucosa.
T1	The tumor is limited the inner layer of bowel
T2	The tumor extends to muscle layer
T3	The tumor invades through the muscle layers but is still inside the serosa
T4a	The tumor invades through the serosa
T4b	The tumor extends to the nearby organs
Node (lymph node)	
N0	No tumor in any lymph nodes
N1a	Cancer cells are in 1 nearby lymph node
N1b	Cancer cells are in 2–3 nearby lymph nodes
N1c	Cancer cells are not in the nearby lymph nodes but have spread into the nearby tissue
N2a	Cancer cells are in 4–6 nearby lymph nodes
N2b	Cancer cells are in more than 7 nearby lymph nodes
Metastasis	
M0	The tumor has not spread to other organs
M1a	The tumor has spread to 1 distant organ
M1b	The tumor has spread to 2 or more distant organs
M1c	The tumor has spread to the peritoneum

Surgical-pathology quality variables also influence prognostic accuracy. Lymph node assessment (lymph node yield) is central because nodal metastasis is a major determinant of adjuvant therapy. Margin assessment is particularly important in rectal cancer (circumferential resection margin), where involved margins are linked to higher local recurrence risk and poorer outcomes [19].

Novel and emerging biomarker

Molecular biomarkers increasingly complement morphology for prognostication and therapy selection. Microsatellite instability (MSI), which results from mismatch repair (MMR) deficiency, is both a Lynch syndrome screening tool and a prognostic marker; MSI-high

tumors are often associated with a more favorable prognosis in early-stage CRC. KRAS (Kirsten rat sarcoma viral oncogene homolog) and BRAF (B-Raf proto-oncogene, serine/threonine kinase) mutations are established predictive biomarkers for resistance to anti-EGFR (epidermal growth factor receptor) therapy in metastatic CRC, while their pure prognostic role in non-metastatic disease is more variable and context-dependent, including interactions with MSI status [19, 22].

Besides, loss of CDX2 (caudal type homeobox 2) expression, HER2 (human epidermal growth factor receptor 2) amplification, enzymatic biomarkers, ctDNA (circulating tumor deoxyribonucleic acid), circulating tumor cells (CTCs), exosomes, or tumor mutation burden (TMB) have also been reported as prognostic biomarkers in CRC, although it is not uniformly adopted across guidelines [19, 22].

Recent work further extends tumor microenvironment (TME) assessment into integrated multimodal signatures intended to improve prognosis prediction and stratify chemotherapy benefit beyond TNM staging alone [20].

Staging and reporting challenges in colorectal cancer prognosis

Despite widespread adoption of structured datasets and TNM staging, multiple steps in CRC prognostication remain interpretive and prone to disagreement, particularly at the interface of “gray-zone” histology and evolving guideline definitions [25, 26]. International survey data show that even among practicing pathologists, recommended staging criteria may be applied inconsistently in specific scenarios, undermining inter-institutional comparability and potentially affecting downstream treatment choice [25].

For instance, although accurate assessment of serosal/peritoneal involvement is crucial because pT4 status drives peritoneal metastasis risk and adjuvant therapy decisions, its evaluation shows only moderate interobserver agreement for distinguishing pT3 from pT4a due to subtle, interpretation-dependent distinctions near the peritoneal surface [27]. In another study, pathologists assigned tumors with inflammatory involvement of the serosa to either the pT3 (49%) or pT4a (51%) stage in nearly equal proportions, independent of experience level, reporting practices, or geographic location. In the same study, authors also discovered that the staging guidelines are not consistently followed by many practicing

pathologists in specific contexts, especially in the assessment of advanced T stages. Notably, in the neoadjuvant context, only 65% based on T-stage classification and margin evaluation on the amount of viable tumor present [25]. In another international survey of CRC pathology reporting practice, tumor budding was never reported by 28.7% of responders, and multiple TNM editions were still in contemporaneous use at the time of survey distribution [26].

Additionally, even when criteria are well-defined, visual assessment and reporting are time-consuming, laborious, and prone to variability across observers as well as within individual observers, particularly for features that require careful scanning of large areas, such as tumor budding, lymphovascular invasion [3, 4]. Workforce constraints amplify these challenges: reports note that a substantial number of elements are expected, workload is increasing, while pathologists are in short supply [28].

In this context, DL-based computational pathology offers a natural bridge: it can leverage whole slide images to produce reproducible, scalable measurements of prognostic morphology and microenvironmental patterns, and can potentially integrate these with molecular and clinical features to improve individualized prognosis prediction [20, 29].

3.2 Deep learning in CRC outcome prediction

As discussed in the previous section, CRC exhibits substantial heterogeneity in its histological features. The current pathological staging system is limited to a set of categorical variables that may not reflect the full spectrum of biology. The staging and pathological reporting are facing several challenges. The rich information contained in histopathological slides is not fully captured by traditional assessment. This gap has motivated a growing interest in computational approaches, particularly AI and DL. These tools can support human experts by detecting additional morphological patterns and revealing prognostic information beyond routine staging criteria.

3.2.1 General trend and achievements

Over the past decade, AI and digital pathology have contributed to the transformation in histopathological analysis. The digitized WSIs have enabled computational analysis to advance in both the quantity and accuracy of histomorphological assessment [30].

Deep learning, a branch of AI that uses multi-layered neural networks to learn visual patterns directly from data, has shown particular promise [29]. These models have demonstrated remarkable accuracy in several diagnostic tasks, including cancer detection, tissue classification, and biomarker quantification across multiple organ systems [30, 31]. For instance, in a recent systematic review and meta-analysis of AI in digital pathology, McGenity et al. reported a mean sensitivity of 96.3% and specificity of 93.3% across diverse diagnostic applications [32]. Despite these very promising results, clinical adoption remains limited [29, 31, 32].

CRC has become one of the most actively studied cancer types for DL-based histopathology analysis. Deep learning has been applied to CRC across a broad range of tasks, such as diagnosis, classification, survival prediction, metastasis prediction, or tumor microenvironment investigation [28]. For instance, Schiele et al. (2021) developed a DL classifier using binary histologic tumor images to predict distant metastasis, demonstrating that tumor H&E images contain predictive information beyond what conventional pathological assessment can extract [33]. Foersch et al. (2023) built a multistain deep learning model that determines an AI-based immunoscore (AImmunoscore) using H&E, CD3, and CD8-stained slides from more than 1,000 CRC patients. This AImmunoscore outperformed other clinical, molecular, and immune cell-based parameters in CRC prognostication, and could also predict response to neoadjuvant therapy in rectal cancer patients [34].

3.2.2 Deep learning in CRC survival prediction

Survival prediction is one of the most clinically significant applications of DL in CRC. Accurate prognostication can directly inform treatment decisions, such as whether a patient should receive adjuvant chemotherapy. The field has evolved considerably over the past several years, progressing through several methodological phases.

Bychkov et al. (2018) published one of the earliest studies that combined convolutional and recurrent neural network architectures (Long Short-Term Memory (LSTM)) to predict CRC outcome directly from tissue microarray images of 420 patients. The model performance outperformed visual histological assessment by expert pathologists in classifying patients into low- and high-risk groups [11]. Kather et al. (2019) trained a convolutional neural network

(CNN) to decompose CRC tissue into nine different classes (CRC epithelium, cancer-associated stroma, adipose tissue, debris, lymphocytes, mucus, smooth muscle, normal colon mucosa, and background) in a first large multicenter study of 862 H&E slides from 500 patients. They introduced a "deep stroma score" and demonstrated that the relative proportion of tumor-associated stroma was an independent prognostic factor for overall survival (OS), disease – specific survival (DSS), and relapse-free survival (RFS) [35].

After that, several studies have been conducted using different advanced approaches. These range from combining manual annotation with CNN-based models [33, 36, 37] to fully end-to-end models [4]. They also developed different architectures, ranging from patch-based CNNs to weakly supervised MIL models. Patch-based CNNs analyze each tile independently before aggregating them into patient-level predictions [38] , whereas weakly supervised MIL models learn from whole-slide labels (e.g., survival data) without requiring per-patch annotations, treating each slide as a bag and tiles as instances [12]. Attention-based MIL helps identify the most prognostic patches [39].

Most recently, foundation models, large-scale models pretrained on massive datasets using SSL, have been applied to CRC survival prediction. One such model was trained on over 130 million patches from more than 104,000 WSIs and then fine-tuned to predict survival outcomes in gastrointestinal cancers, including identifying patients who would benefit from adjuvant chemotherapy [14]. Another recent approach, the Prognostic Representation of Integrated Spatial Morphology (PRISM) model, is a morphology-aware prognostic framework that explicitly encodes 13 distinct CRC tissue morphologies alongside generic foundation-model features and uses an MIL architecture to aggregate patch-level information [16].

Multimodal and integrative approaches.

Another important direction is combining histopathology images with other data types. One interpretable, weakly supervised multimodal framework fused WSIs with RNA-sequencing, copy number variation, and mutation data from 5,720 patients across 14 cancer types, achieving improved risk stratification in 9 of 14 cancers [40]. Another study integrated morphological features from WSIs with clinical variables and mutation signatures to improve colon adenocarcinoma stratification, particularly for moderate-risk patients who are often

the most difficult to classify [41]. A further line of work showed that combining genomics and histopathological features can outperform single-modality models for colon cancer stage classification and survival prediction [42]. More recently, a multimodal tumour microenvironment signature combining pathomics, immune scores, and collagen signatures has been proposed for predicting prognosis and chemotherapy benefit in CRC [20].

Remaining challenges

Despite these achievements, several significant challenges remain. Firstly, most studies still lacked multicenter external validation or had very few samples. For example, one model was trained on only 74 cases out of a total of 129 patients in a single cohort [36], and another used just 128 patients from a single institution [33]. Sun et al. (2022) mentioned that models may rely on cohort-specific characteristics, leading to optimistic results that cannot be generalized [37]. The second challenge is the variability in technique. Standardizing processes, such as image preprocessing or building robust models, needs to be addressed to reduce that variation [30]. Thirdly, critically for clinical adoption, the lack of model interpretability remains a major barrier. Most of DL-based models are lacking interpretation because their data-driven training makes the internal decision processes difficult for humans to understand, while pathologists and oncologists need to understand AI-based predictions before they can apply them in clinical decision-making [8, 29].

3.2.3 Interpretability of the AI models

The Lack of Model Transparency and Efforts to Interpret Models

Historically, most DL studies in CRC survival prediction concentrated primarily on maximizing predictive performance, such as achieving higher AUCs (Area under the curve), higher hazard ratios (HR), and better Concordance indices (C-indices), with relatively little attention paid to model interpretability or relying barely on latent features [11, 37]. While these models showed statistical significance in survival prediction, the absence of meaningful, interpretable features limited their practical value [30].

Furthermore, many supervised approaches required extensive annotations from expert pathologists. For instance, pixel-level or region-level labels identifying different tissue types are time-consuming, expensive, and difficult to scale [29, 35]. This annotation burden not only restricted the size of training datasets but also introduced a degree of subjectivity, since different pathologists may annotate the same tissue regions differently [10, 29].

Recognizing these limitations, a growing number of recent studies have moved toward building self-supervised or weakly supervised models that can learn meaningful representations from unlabeled data [14, 15, 43], while also making efforts to interpret the features driving model predictions.

Some common interpretation methods that have been proposed in recent studies for making deep learning models more transparent in this field are listed here:

- **Gradient-weighted Class Activation Mapping (Grad-CAM):** Grad-CAM generates heatmaps that highlight the image regions contributing most to the model's prediction by utilizing the gradient information from the last layer of CNN. It is one of the most widely used visualization methods in histopathology and has been applied in multiple CRC prognostic studies to localize prognostically relevant tissue areas [9, 44, 45].
- **Shapley Additive Explanations (SHAP):** Based on game theory, SHAP values provide a measure of each feature's contribution to a given prediction, enabling researchers to rank and compare feature importance. In some recent studies, authors applied SHAP to interpret their CRC prognostic [20, 44].
- **Attention maps:** In attention-based MIL models, the learned attention weights assigned to different image patches serve as a natural form of interpretability, indicating which WSI regions the model considers most important for its prediction [4, 46].

These methods provide valuable insights into how the model behaves. Grad-CAM and attention heatmaps illustrate which areas of the image have the strongest influence on a prediction, so they are useful to see where the model is “looking” in the tissue. SHAP, on the other hand, shows how much each feature contributes to the prediction

However, in computational pathology, they still have limitations. They can show which regions or features matter, but they usually do not provide clear information about what specific morphological patterns or tissue structures the model has actually learned. In addition, the outputs are often hard to turn into clean, well-defined morphological features that pathologists can easily relate to their own concepts and use in everyday clinical interpretation.

Common Model-Driven Features

Despite the interpretability challenges, various interpretation approaches have converged on several recurrent morphological themes as prognostically relevant features in CRC:

- **Tumor-stroma composition:** The proportion and spatial organization of stroma relative to tumor tissue has been consistently identified as a key prognostic factor. A deep stroma score derived from histology has been shown to serve as an independent prognostic marker [35], and automated tumor–stroma ratio (TSR) assessment has also emerged as a significant predictor of survival [47, 48].

- **Immune cell infiltration:** The density and spatial distribution of immune cells, particularly TILs, have been repeatedly linked to CRC prognosis. Quantifying CD3+ and CD8+ T cell density in the stroma provides a prognostic stroma–immune score [49], and high TIL density near the invasive margin correlates with improved progression-free survival [50]. An AI-based immunoscore has also been shown to outperform traditional immune cell–based parameters [34].

- **Tumor differentiation and nuclear morphology:** Poorly differentiated tumor regions, nuclear pleomorphism, and abnormal glandular structures have been associated with worse outcomes. Quantitative analysis of cellular morphological features within the tumor region has demonstrated prognostic value, with high cellularity, large tumor nuclei, and abnormal architectural structures characteristic of high-risk patients [51][41].

- **Tumor microenvironment features:** Features related to the microenvironment, such as adipose tissue adjacent to the tumor and fibrosis, have been highlighted by several models. Clusters of tumor cells close to adipose tissue and intermediate or mature

desmoplastic stromal reactions are strongly associated with high-risk predictions [10]. Adipocyte accumulation and fibrous tissue deposition are key components of prognostic pathomics signatures [44].

Beyond Heatmaps: Deeper Interpretation

While most studies have relied on heatmap-based visualizations (Grad-CAM, attention maps) to provide a general sense of which slide regions drive predictions, very few have attempted to go deeper, such as systematically categorizing and clustering the morphological patterns that underlie model decisions. The heatmaps can offer a starting point, but are not sufficient for building the deep trust needed for clinical adoption, especially in cancer.

The most notable effort in this direction is the work of Wulczyn et al. (2021), published in *npj Digital Medicine*. After developing a deep learning system for predicting DSS in stage II and III CRC, they went beyond standard interpretation methods to investigate what the model learned. They first showed that standard Tumor, Node - categories, and grading explained only a small fraction of the variance in model scores (R^2 = 18%). The authors then generated human-interpretable histologic features by clustering embeddings from an external deep learning-based image-similarity model and showed that these cluster-derived features explained the majority of the variance (R^2 = 73–80%). They discovered the cluster most strongly associated with high-risk scores was “tumor adjacent to adipose” pattern and could be reliably recognized by human annotators, achieving an accuracy of 87.0–95.5%. This study was significant because it moved beyond simply showing where the model focuses (heatmaps) to identifying what morphological patterns the model uses, and pathologists can review, understand and validate [10].

Sajjad et al. (2025) also incorporated morphological awareness into their PRISM model by training a morphology-informed classifier on 13 distinct CRC tissue morphologies and using the learned embeddings to capture the continuous spectrum of phenotypic variability within tumor architecture. This approach integrated morphology-aware features with generic foundation model features, providing both improved performance and richer biological interpretability [16].

However, a review of the current literature reveals that few studies have gone beyond the level of heatmap visualization or single-step clustering. Specifically, the systematic multi-step strategy of (1) generating attention heatmaps to identify high-attention regions, (2) analyzing the spatial distribution of top-scoring tiles across the WSI, (3) clustering these top tiles into distinct morphological groups, and (4) identifying the most informative clusters for clinical interpretation represents a deeper layer of analysis that is largely absent in the existing literature. Wulczyn et al.'s clustering approach comes closest, but it focused on clustering all tile embeddings globally rather than specifically examining the top attention-weighted tiles [10]. The multi-step interpretive approach (heatmap, spatial distribution, morphological clustering of the most informative tiles) could provide pathologists not only with a risk score, but also with a structured set of visual features and their spatial context that explain the prediction. This multi-layered interpretation strategy aligns with the way pathologists do in practice: they assess the overall architecture of the tissue, identify key regions, and then characterize the specific patterns within those regions before reaching a diagnostic or prognostic conclusion.

The demand for this level of interpretability stems from a fundamental reality in clinical oncology: every prognostic and treatment decision results in consequences for patients. Clinicians need to understand “why” an AI model assigns a particular risk level before they can responsibly incorporate that information into clinical practice. By moving toward structured morphological cluster identification and characterization, AI-based prognostic tools can become more transparent, clinically meaningful, and ultimately more likely to be adopted by the pathology community.

3.2.4 Summary of key studies

Recent studies have explored a wide range of deep learning approaches for survival prediction in colorectal cancer. Models ranged from classical convolutional neural networks (e.g., VGG-16) to transformer-based and multiple instance learning frameworks (e.g., TransMIL, DeepDisMISL), as well as foundation models such as DINOPath. Taken together, the literature shows a gradual shift from supervised models focused mainly on predictive performance toward more interpretable and clinically oriented approaches, with increasing

attention to explainability; however, systematic evaluation of these explanations remains limited. A summary of representative studies is provided in Table 3.2.

Table 3.2: Summary of the most related DL-based survival models for CRC outcome prediction

First author(year)	Dataset	Prediction strategy	Model performance
Bychkov (2018) [11]	420 patients (1 cohort, TMAs): Helsinki University Central Hospital (Finland)	VGG-16 + recurrent neural network (Long Short-Term Memory)	HR 2.3 (CI 95% 1.79–3.03); AUC 0.69
Wulczyn (2021) [10]	5,629 patients (1 cohort, 3 splits): Medical University of Graz Biobank, stage II-III CRC, resection cases 1984–2013 (Austria): Training + tuning set: 3,652 cases Validation set 1 (1984-2007): 1,239 cases Validation set 2 (2008-2013): 738 cases	Weakly supervised deep learning system (DLS)	AUC 0.69 – 0.70
Li (2022) [46]	1,713 patients (2 cohorts): MCO (Australia): 1,184 patients TCGA-COAD/READ (USA): 529 patients	Distribution based multiple-instance survival learning algorithm (DeepDisMISL)	C-index = 0.638
Zhou (2023) [41]	623 patients (3 cohorts): TCGA-COAD (USA): 336 patients WSU-COAD (USA): 123 patients TCGA-READ (USA): 164 patients	InceptionV3: image feature extraction Integrative multimodal model (image features, clinical variables, genomic variables)	AUC = 0.81 ± 0.08
Jiang (2024) [4]	4,428 patients (4 cohorts): DACHS (Germany): 2271 patients MCO (Australia): 1395 patients TCGA-CRC (USA): 565 patients DUSSEL (Germany): 197 patients	Self-Supervised Learning (SSL) + Attention-based Multiple Instance Learning (attMIL)	HR 4.50 (internal test set) HR 2.23 – 3.08 (three external validation sets)
Wei (2024) [9]	1,264 patients (2 cohorts): PLCO (USA): 640 patients TCGA-CRC (USA): 624 patients	U-Net TransMIL	C-index = 0.748
Wang (2025) [14]	4,213 patients (7 cohorts†): 1,619 patients across 4 gastric/esophageal cohorts (multi-national); 2,594 CRC patients across 3 CRC cohorts:	Foundation model: DINOPath	C-index = 0.714 – 0.757 for colorectal cancer (DSS)

First author(year)	Dataset	Prediction strategy	Model performance
	PLCO (USA): 465 patients (training), 196 (validation). TCGA-CRC (USA): 587 patients MCO (Australia): 1346 patients		
Sajjad (2025) [16]	424 patients (1 cohort): Alliance (CALGB 89803) cohort, stage III CRC, randomised phase III trial (USA, multi-centre)	Prognostic Representation of Integrated Spatial Morphology (PRISM) model	AUC = 0.70 ± 0.04, accuracy = 68.37% ± 4.75%, HR = 3.34 (95% CI: 2.28-4.90)

Abbreviations: CRC, colorectal cancer; TMA, tissue microarray; TCGA, The Cancer Genome Atlas; COAD, colon adenocarcinoma; READ, rectal adenocarcinoma; MCO, Molecular and Cellular Oncology cohort (Australia); DACHS, Darmkrebs: Chancen der Verhütung durch Screening (Germany); PLCO, Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (USA); DUSSEL, Düsseldorf cohort (Germany); WSU, Wayne State University cohort (USA); HR, hazard ratio; AUC, area under the curve; CI, confidence interval.

3.3 Overview of the computational workflow

This section provides a brief overview of the computational pipeline used to generate the model outputs analysed in this thesis. The model development was conducted in a previous study of Laurila [52]; therefore, only the essential steps required to understand the origin of the embeddings, risk predictions and attention maps are described here.

3.3.1 Whole slide image processing

Whole slide images (WSIs) are digitized histopathology slides scanned at high resolution, often exceeding 100,000 × 100,000 pixels. Due to their enormous size, WSIs cannot be processed directly by standard deep learning models and require a series of preprocessing steps before analysis [53].

Tissue segmentation:

The first step in WSI preprocessing is tissue segmentation, which separates the actual tissue regions from the glass slide background. This step ensures that only diagnostically relevant regions are retained for downstream analysis, discarding the empty background that constitutes a large portion of the scanned image [53, 54].

Color normalization:

Histopathology slides stained with H&E can exhibit significant color variation due to differences in staining protocols, reagent concentrations, and scanner hardware across

laboratories. Color (stain) normalization addresses this variability to ensure consistent appearance across slides [55]. The Macenko method is one of the most widely adopted approaches: it decomposes the image into its stain components using singular value decomposition (SVD) in the optical density color space, then rescales these components to match a reference image. By standardizing color representation, normalization enables more reliable and reproducible computational analysis [56].

Patch extraction (tiling):

WSIs are divided into smaller, non-overlapping patches (tiles), typically sized 256 × 256 pixels at a chosen magnification level, because their original sizes are too large to be fed into neural networks as a whole. This process is called tiling or patch extraction. The resulting collection of patches from a single WSI can number in the hundreds to thousands, depending on the required magnification, and each patch serves as an individual input unit for feature extraction in subsequent computational steps [13, 53]

3.3.2 Foundation model embeddings

Foundation models are large-scale neural networks pretrained on vast amounts of data using SSL, designed to learn general-purpose feature representations that can be transferred to a variety of downstream tasks. In computational pathology, these models are typically trained on millions of histopathology image patches extracted from thousands of WSIs [43].

These models capture morphological patterns such as cellular structure and tissue organization and encode them into compact numerical representations (embeddings). Once pretrained, foundation models can be applied to many different tasks, including tumor detection, classification, segmentation, with minimal or no additional labeled data, substantially reducing the annotation burden that has historically limited computational pathology [15].

Several large-scale foundation models trained specifically for pathology have recently been introduced, such as UNI and UNI2-h, which was developed Mahmood Lab [15], or DINOPath, a model that was pretrained with over 130 million patches derived from 25 different organs [14]. In this thesis, UNI2-h is used for feature extraction because it provides high-quality

representations and has shown strong performance across various histopathology tasks [15, 57].

3.3.3 Multiple instance learning framework

Multiple instance learning (MIL) is a weakly supervised learning framework. In MIL, training data are organized into groups called "bags", each containing multiple unlabeled "instances". In the histopathological context, “bag” equals WSI, and “instances” equals image patches extracted from the WSI through preprocessing. A bag is labeled positive if at least one of its instances is positive, and negative only if all instances are negative. The learning objective is to predict labels for new, unseen bags based on this bag-level supervision [58, 59].

In computational pathology, MIL is particularly well-suited because exhaustive pixel-level or patch-level annotation of gigapixel WSIs is impractical. Only a slide-level diagnosis label (e.g., tumor-positive or tumor-negative) is needed for training. Modern MIL approaches, such as attention-based MIL and CLAM, use attention mechanisms to assign an importance score to each patch, allowing the model to learn which tissue regions are most relevant for the slide-level prediction. This enables accurate classification, including cancer detection, subtyping, and biomarker prediction, without requiring detailed spatial annotations [58, 59].

Among these, CLAM, proposed by Lu et al. (2021), was selected as the MIL framework in this thesis due to several practical strengths. CLAM combines attention-based pooling with instance-level clustering to refine the feature space, which not only improves classification accuracy but also produces interpretable attention heatmaps that highlight diagnostically relevant tissue regions. Its open-source implementation further provides an end-to-end pipeline covering WSI segmentation, feature extraction, and model training, making it a practical and reproducible choice for weakly supervised computational pathology studies [13].

3.3.4 K-means clustering algorithm

K-means is an unsupervised machine learning algorithm used to partition data into a predefined number of groups (K clusters) based on similarity. Despite this algorithm has been proposed for more than 50 years, it is considered the most popular clustering method. K-

means clustering is simple, straightforward implementation, computational efficiency, and strong performance in practice [60].

In computational pathology, K-means clustering can be used to group morphologically similar tissue patches based on their feature embeddings, thereby identifying recurring tissue patterns or phenotypes within a dataset without the need for manual annotation. A key consideration is the choice of K, which is often determined using the "elbow method", plotting clustering error against increasing values of K and selecting the point where additional clusters yield diminishing improvement [61]. When the elbow method does not yield a clear inflection point, the silhouette method can be used as a complementary approach. The silhouette score measures how well each data point fits within its assigned cluster relative to neighboring clusters, producing a value between −1 and +1, where higher values indicate better-defined clusters [62]. Besides, Jain et al. (2010) mentioned that finding the optimal k value is not an easy task [60].

3.3.5 Forward stepwise regression

Forward stepwise regression is a variable selection method used to build a parsimonious regression model from a large set of candidate predictor variables. This method has been widely used in medical studies. The procedure begins with an empty model containing no predictors. At each step, every remaining variable is tested for inclusion, and the one that most improves the model fit, which is measured by criteria such as the lowest p-value, highest increase in R^2, or greatest reduction in Akaike Information Criterion (AIC), is added. This process continues iteratively until no remaining variable provides a statistically significant improvement to the model [63].

Forward stepwise regression is particularly useful when the number of candidate variables is large relative to the sample size, as it avoids the computational expense of evaluating all possible subsets, and it is simple to apply. In clinical and biomarker research, it is commonly applied to identify a manageable set of predictors, from a pool of clinical, molecular, or imaging-derived features, that are most strongly associated with the outcome of interest. However, it should be noted that forward selection can sometimes miss important variables, and model results should be interpreted with appropriate caution [63].

4 Methods

4.1 Data Collection and Cohort Description

A retrospective large dataset of H&E-stained WSIs from patients diagnosed with CRC was assembled. All samples were collected from surgical resections at Helsinki University Hospital (HUS) between 1982 and 2009 (collected in collaboration with Professor Caj Haglund). The histopathology slides were processed using standard protocols and then digitized at high resolution (~0.24 µm per pixel, equivalent to 40× magnification) with a 3DHISTECH Pannoramic 250 scanner.

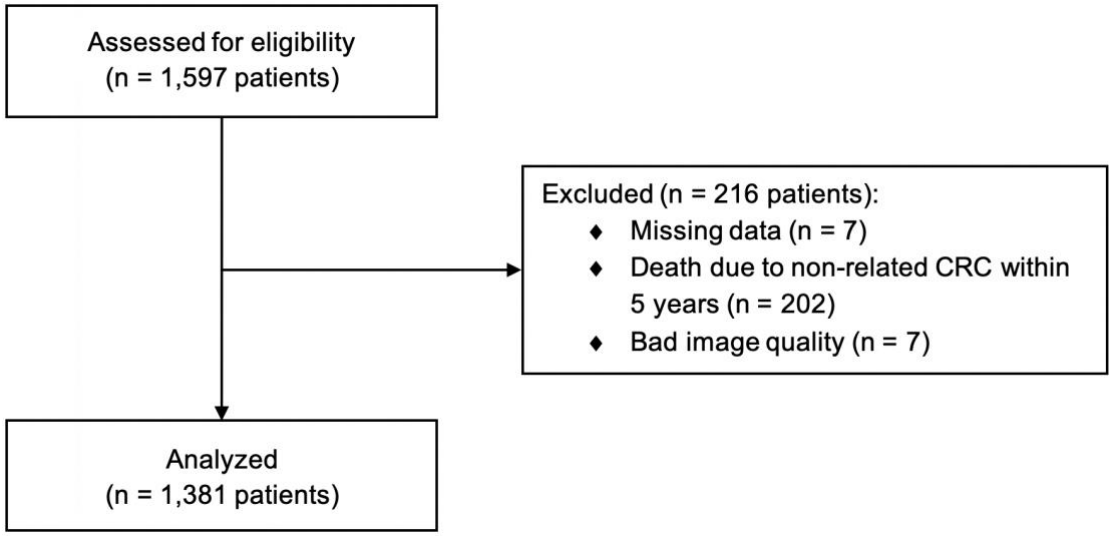


Figure 4.1: Flow diagram of patients and WSIs selection

Clinical data and follow-up information were obtained for each patient from hospital records and national registries (Finnish Digital and Population Data Services Agency and Statistics Finland). The primary outcome of interest was 5-year survival status, approximating DSS, defined as the time to death due to CRC within five years of surgery [64]. Patients who died not because of CRC within five years were excluded to focus on cancer-specific outcomes. Patients who survived at the 5-year end were considered “alive”, whereas those who died due to CRC within five years were considered “dead”.

Figure 4.1 illustrates the flow to select patients, and finally, 1,381 patients with 1,407 WSIs were included, as some patients had multiple WSIs in this dataset. This cohort is a consecutive series of CRC cases encompassing a wide range of tumor stages and histological grades. At the 5-year follow-up, 881 patients (63.8%) were alive, and 500 (36.2%)

had died from CRC. The sex distribution was even (50.1% male), and the median age at surgery was 67.4 years.

All patient materials were used in accordance with ethical approvals and regulations. The study was approved by the HUS Ethics Committee (approval HUS/1223/2021) and conducted under a research permit from HUS (HUS/74/2023), both held by Prof. Caj Haglund. Authorization to use archived tissue samples was granted by the Finnish Medicines Agency (FIMEA/2021/006901). All tissue specimens and clinical records were assigned unique identifiers prior to analysis, and no identifiable patient information was available to the researchers at any stage.

4.2 Model Overview

This section provides a brief overview of the image preprocessing steps (Figure 4.2) and the deep learning pipeline (Figure 4.3) used in this study. The primary goal was not to develop a new prediction model but to analyze and interpret the outputs generated by an existing model. Detailed descriptions of the model architecture and training procedures are available in the previously published thesis by Amanda Laurila [52].

4.2.1 Image preprocessing

Tissue segmentation and slide splitting

Binary tissue masks (tissue vs. background) were generated using an existing segmentation algorithm and applied to crop individual tissue regions and remove background areas [52]. A single WSI could contain multiple disconnected tissue fragments (e.g., multiple tissue pieces within the same paraffin block), each fragment was treated as an independent tissue section and saved as a separate image. In rare cases where the automatically generated masks were inaccurate, manual corrections were applied to ensure proper tissue representation. As a result, a single WSI could produce multiple tissue section images, leading to a total of 1,950 tissue section images being obtained for analysis.

Stain normalization

The color intensity and tone of the H&E stain varied because the slides were collected across different years and staining batches. Stain normalization was performed using Macenko's

method to reduce this variability. This normalization made the purple and pink hues of H&E more consistent across the entire dataset, so that color differences would not confuse the model.

Patch extraction (tiling)

In the final preprocessing step, each tissue section image was divided into non-overlapping patches of 256 × 256 pixels (px) at 40X magnification for input to the deep learning model. On average, each WSI yielded tens of thousands of such tiles. The patch coordinates were recorded, but only the patch images were used during model training.

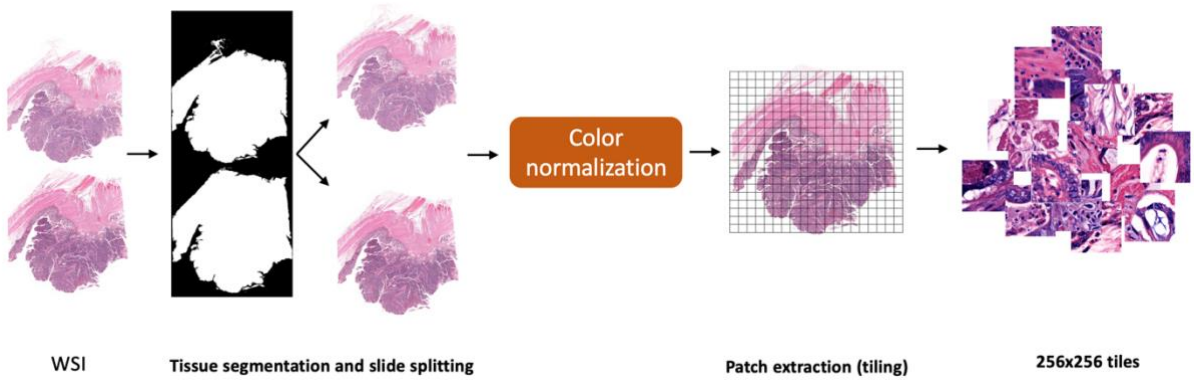


Figure 4.2: Preprocessing steps were performed to prepare the WSIs before feeding into the DL model.

4.2.2 UNI2-h-CLAM survival prediction model

The foundation model (UNI2-h) was used to extract feature representations, which were utilized for the morphological pattern exploration task, from each image patch. A multiple instance learning framework (CLAM) was used to aggregate patch-level feature embeddings into an image-level representation via an attention mechanism. The model produced a predicted survival probability at the image level, which serves as the basis for downstream analyses in this study, including attention-based visualization. Detailed descriptions of the model architecture and training procedures are available in the thesis by Amanda Lauria [52]

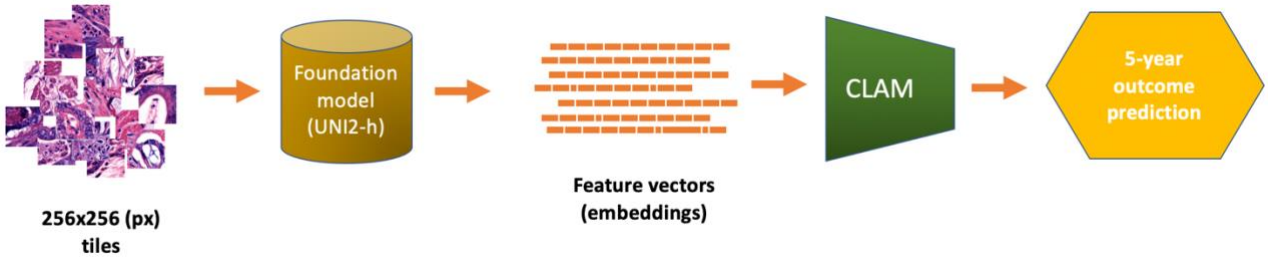


Figure 4.3: Overview of UNI2-h-CLAM survival model pipeline

4.2.3 Statistical analysis

All computational analyses were performed in Python 3.8, using standard libraries for data handling and evaluation: NumPy and pandas for data manipulation, scikit-learn for model training and classification metrics, and Matplotlib for plotting. For survival analysis, the lifelines package was utilized, which provided functions for Kaplan-Meier survival curves, log-rank tests, and Cox proportional-hazards regression.

Image-level vs. patient-level predictions

Image-level predictions were aggregated to obtain a patient-level prediction because some patients had multiple WSIs or many tissue fragments extracted from one slide.

For patients with multiple images, predicted probabilities were averaged to produce a single patient-level score. A threshold of 0.5 was then applied to make a binary prediction for survival: if the patient's risk score was ≥ 0.5, the patient was predicted to be "dead", whereas a score < 0.5 predicted the patient would be alive at least 5 years.

Model performance evaluation

The model's predictive performance was evaluated using the held-out test sets from the 5-fold cross-validation. Metrics included accuracy, sensitivity, specificity, F1-score, area under the receiver operating characteristic curve (AUC ROC), and concordance index (C-index). F1-score was used to balance precision (positive predictive value) and recall (sensitivity), especially when the outcome groups were not evenly distributed [65]. The C-index was used to assess how well the model ranked patients according to survival risk, which is important for prognostic prediction [66]. Results were summarized as the mean ± standard deviation (SD) across the five folds.

Survival curve analysis

To relate the model's predictions to actual patient survival over time, a Kaplan-Meier survival analysis was conducted. Patients were divided into two groups based on the model's prediction ("dead" and "alive") using the patient-level classification from the model. The log-rank test was computed to statistically assess whether the survival curves of the two predicted groups were significantly different. In addition, the Cox proportional-hazards model was employed to estimate the hazard ratio (HR) between the two groups. HR was reported with 95% confidence intervals (CI) to indicate the strength and uncertainty of this effect.

4.3 Model-Driven Feature Interpretation

After training the deep learning model, a series of post-hoc analyses were performed to interpret what histopathological patterns the model had learned. The goal was to identify which tissue morphologies were being used by the model to distinguish alive from dead patients, and to relate those features back to known pathology. The following subsections describe the steps that were utilized to derive and examine model-driven features and patterns from the WSIs.

4.3.1 Post-hoc morphological pattern analysis

Slide selection for interpretation

First, the images that the model predicted either "dead" or "alive" with high confidence were focused on. Using the test results from the cross-validation, the top confidently predicted (alive or dead probability≥ 0.8) images were taken from each fold. These high-confidence cases from all folds were pooled together to create a set of images for in-depth interpretation.

Attention heatmaps and top tiles

For each selected image, an attention heatmap over the WSI was generated to visualize which areas the model found most important. Then the top 50 patches with the highest attention scores from each slide were extracted. These top tiles represent the image regions that contributed the most to the model's prediction for that image. Spatial distribution maps were conducted to show the locations of these high-attention patches. These maps were examined to determine whether the patches clustered in specific tissue regions (for example,

within the tumor epithelium or at the invasive front) or were more diffusely distributed. A high-attention subset was created by pooling the top patches from slides with high level of confidence and used for subsequent analysis of common morphological features.

4.3.2 Feature clustering

Each of the high-attention tiles in the subset had been passed through the foundation model (UNI2-h) to obtain a feature vector – a numerical representation of the patch's visual features. K-means clustering was then applied to these feature vectors (using the scikit-learn implementation) to group patches that had similar features. The aim was to identify a set of data-driven morphological clusters – patterns of tissue architecture or composition that the model might be using in its decision-making.

Determining the number of feature clusters: Different numbers of clusters k were experimented with, and the silhouette analysis was used to guide the choice of k because elbow method did not yield a clear inflection point. The silhouette scores were calculated for a range of cluster counts (Figure 4.4). Overall, the silhouette values were relatively low, which is appropriate given the high complexity and overlap in histology image features. It showed that increasing the number of clusters gradually improved the score, whereas using too few clusters caused very dissimilar patches to be forced into the same cluster. The low values of k (less than 75) were excluded because of these reasons: firstly, most of these k values have low silhouette values, except k = 15; secondly, they yielded overly broad clusters that mixed different tissue patterns, reducing interpretability. Instead, a relatively high number of clusters were chosen in order to capture more fine-grained distinctions between patch groups. Based on the silhouette analysis (Figure 4.4), the values of k explored were: 75, 85, 100, 105, 125, and 135 for further interpretation, accepting that the clusters would be somewhat overlapping given the complex feature space.

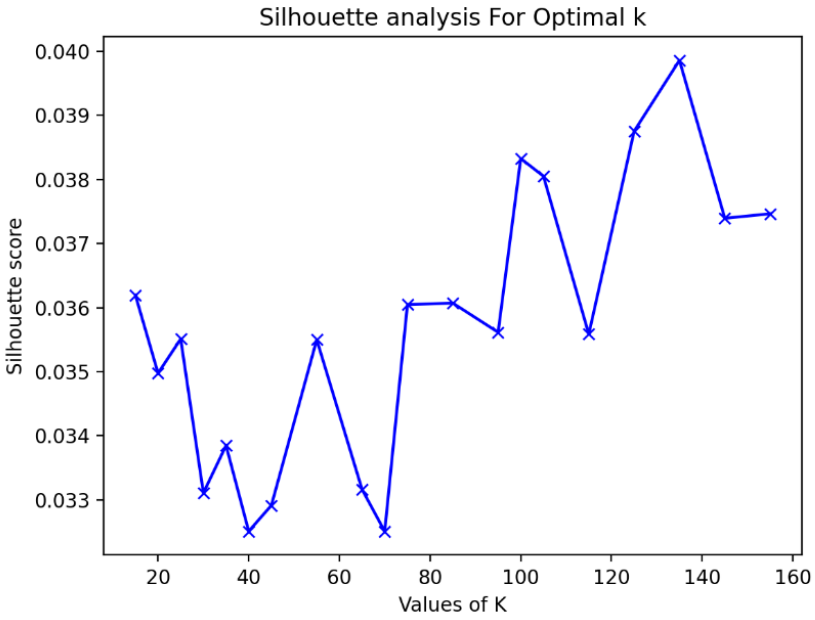


Figure 4.4: Silhouette scores for k-means clustering across a range of k values

Identifying prognostic clusters: After clustering the image tiles, each cluster was treated as a potential morphological feature. To examine how these features were related to patient survival, the analysis was carried out at the image level rather than at the level of individual tiles. For each image, the tiles were assigned to clusters, and the number of tiles in each cluster was counted. In this way, each slide was represented by the distribution of tissue patterns in its most informative regions. These cluster distributions were then compared between patients with different outcomes. Because the predicted probability of survival was the complement of the predicted probability of death, the forward stepwise selection was performed using the dead probability as the outcome measure. This allowed the clusters that best distinguished between the survival groups to be identified. Based on this procedure, the ten most informative clusters were selected and considered the main histological patterns contributing to the model's survival prediction.

Cluster interpretation with example tiles: To interpret the histopathological meaning of each top cluster, representative patches were selected from the corresponding image tiles. For each of the top 10 clusters, the 15 tiles closest to the cluster centroid in feature space were retrieved. Patches were sampled from different patients' images to ensure that the cluster pattern reflected the cohort as a whole rather than a single outlier case. The patches were then reviewed visually to identify common themes. By linking each important cluster to

representative histology images, the key tissue morphologies associated with the model's prognostic predictions could be characterized. Through this cluster-wise inspection, insight was gained into the histologic patterns that were used by the model to distinguish between the alive and dead groups.

4.3.3 Dimensionality reduction and visualization

To complement the clustering analysis, the high-dimensional patch feature space was visualized in two dimensions. Non-linear dimensionality reduction methods, t-distributed Stochastic Neighbor Embedding (t-SNE), was used to project the patch feature vectors onto a 2D plane. This method was chosen because it is well-suited for visualizing complex high-dimensional data and for preserving local structure among similar patches. In this way, scatter plots were created in which each point represented a patch while maintaining local similarity [67].

For visualization, the tile embeddings were displayed using two coloring schemes. In the first scheme, the tiles were colored according to their K-means cluster assignment, which highlighted groups of tiles with similar feature representations. In the second scheme, the tiles were colored according to their originating WSIs or slide ID. This dual visualization strategy was used to assess whether the clustering mainly captured shared morphological patterns across slides or whether the tiles were grouped primarily by slide-specific characteristics.

4.3.4 Error analysis and model interpretation

Finally, a qualitative error analysis was conducted to examine the model's failure cases. Misclassified test cases with high prediction confidence were reviewed to understand where the model had gone wrong. For each case, the WSI, attention heatmap, and top 50 patches were inspected, and the corresponding tissue structures were reviewed. This analysis was used to identify the patterns most often associated with misclassification and to better understand the model's limitations.

5 Results

5.1 Model performance evaluation

The performance of the proposed model was evaluated at both the image level and the patient level. Image-level predictions generated by the CLAM model were first assessed and demonstrated stable performance across five cross-validation folds. Overall image-level performance was obtained by averaging the results across all folds. Applying stain color normalization (Macenko's method) resulted in only modest performance gains (Table 5.1).

Table 5.1: Performance of the CLAM model trained at the image level with and without color normalization. Metric values are reported as mean ± SD, averaged across five cross-validation folds.

Metric	Without color normalization	With color normalization
AUC	0.74 ± 0.018	0.75 ± 0.020
Accuracy	0.68 ± 0.007	0.70 ± 0.033
F1 - score	0.48 ± 0.016	0.57 ± 0.048
Sensitivity	0.47 ± 0.202	0.55 ± 0.084
Specificity	0.80 ± 0.106	0.79 ± 0.078

For patient-level prediction, the image-level outputs from the same patient were averaged. As shown in Table 5.2, this aggregation improved performance across all evaluation metrics compared with image-level prediction. Specifically, patient-level prediction achieved higher accuracy and AUC ROC, indicating better overall classification performance and improved discrimination capability. In addition, F1 score, sensitivity, and specificity also showed similar increasing trends, suggesting more balanced performance and more reliable identification of positive and negative cases at the patient level.

Overall, these results demonstrate that aggregating image-level predictions to the patient level leads to consistently better and more stable performance across multiple metrics.

Table 5.2: Performance of the CLAM model with color normalization preprocessing at image- and patient-level. Metric values are represented as mean ± SD which was averaged across cross-validation folds

Metric	Image-level	Patient-level
AUC	0.75 ± 0.020	0.79 ± 0.034
Accuracy	0.70 ± 0.033	0.73 ± 0.043
F1 - score	0.57 ± 0.048	0.62 ± 0.075
Sensitivity	0.55 ± 0.084	0.62 ± 0.120
Specificity	0.79 ± 0.078	0.80 ± 0.051

To more precisely assess prognostic discrimination, the C-index was calculated for the CLAM model using predicted probabilities as continuous risk scores across the entire follow-up period, instead of relying on the binary 5-year outcome. The CLAM model achieved a concordance of 0.738.

Kaplan–Meier analysis was conducted to evaluate the relevance between model outputs and clinical outcomes. The Kaplan–Meier curve in Figure 5.1 demonstrated clear survival stratification by CLAM predictions. Patients classified as alive at 5 years exhibited significantly higher survival than those classified as dead, with a hazard ratio (HR) of 4.27 (95% CI, 3.22–5.66; log-rank p-value <0.001).

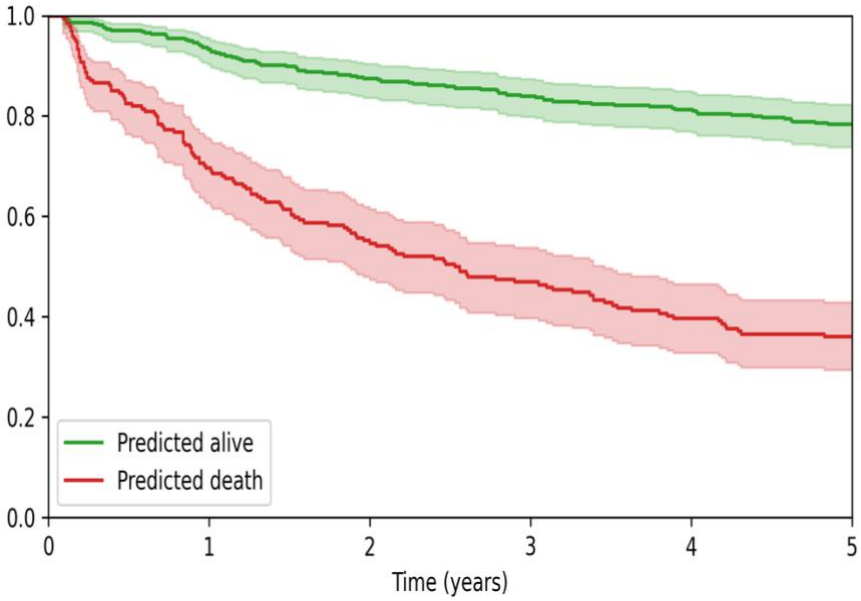


Figure 5.1: Kaplan–Meier survival curves stratified by CLAM–predicted 5-year outcomes.

Alive-predicted patients (green) demonstrated significantly better observed survival compared with those dead-predicted (red). Sample sizes: n = 374 (alive) and n = 194 (dead). Shaded regions indicate 95% confidence intervals

5.2 Model-Driven Feature Representation and Visualization

5.2.1 Top tiles distribution and morphology

This step identifies which image regions are most important for predicting patient outcomes. The heatmaps illustrated the most crucial regions in dark red, the non-association areas in blue, based on attention scores. In most of the images, the model's attention concentrated on tumor areas, invasion regions, or stroma surrounding the tumor, whereas the model skipped the necrosis, mucinous areas, or benign tissue (muscle, fat tissue) (Figure 5.2). The spatial distribution maps of the top 50 tiles of each image showed the concentration of these tiles. The top dead-predicted tiles were mainly located in invasion zones, including the serosa and surrounding fat. In contrast, top alive-predicted tiles were more often located in the tumor epithelium and muscularis propria.

In most of the high probability dead-predicted images, the top 50 tiles showed moderate- to high-grade nuclear atypia (enlarged, dark, irregular border nuclei, big and prominent nucleoli, coarse, clumped chromatin; or bizarre nuclei), poorly differentiated or solid tumor growth (poorly formed gland, or clusters or tumor cells, or separate individual tumor nucleus), and marked immature (myxoid) or intermediate (keloid) desmoplastic stroma. The qualitative

evaluation of the tumor-stroma ratio suggested that the stroma areas were predominant (estimated >50% of the tile area) in the majority of dead-predicted images. These features are consistent with histological patterns previously associated with aggressive disease and poor prognosis (Figure 5.3). By contrast, most of the top tiles in alive-cases showed preserved gland architecture, low- to moderate-grade nuclei, and frequent inflammatory cells (mostly acute inflammatory cells). These findings are typically associated with better-differentiated tumors and improved outcomes (Figure 5.4)

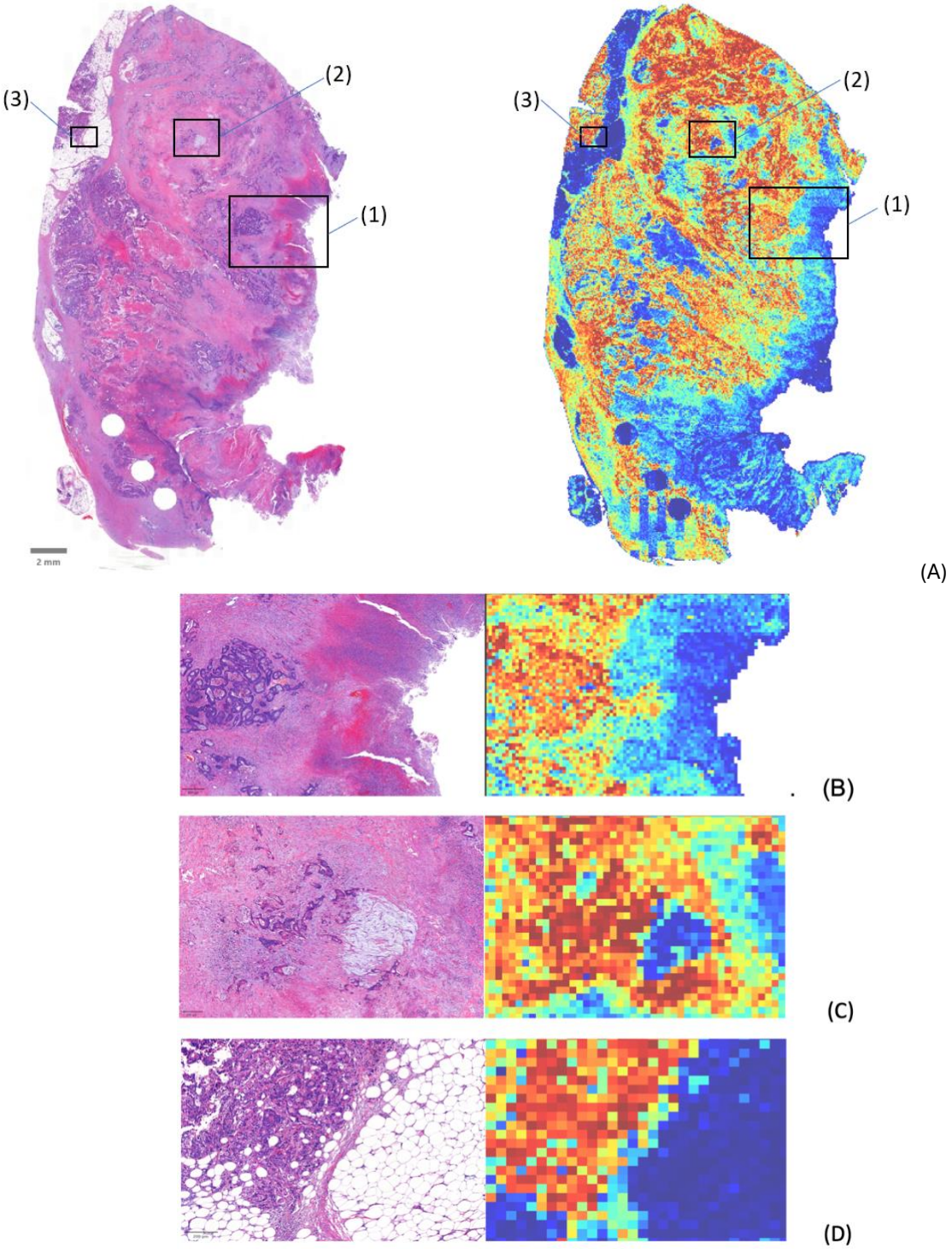
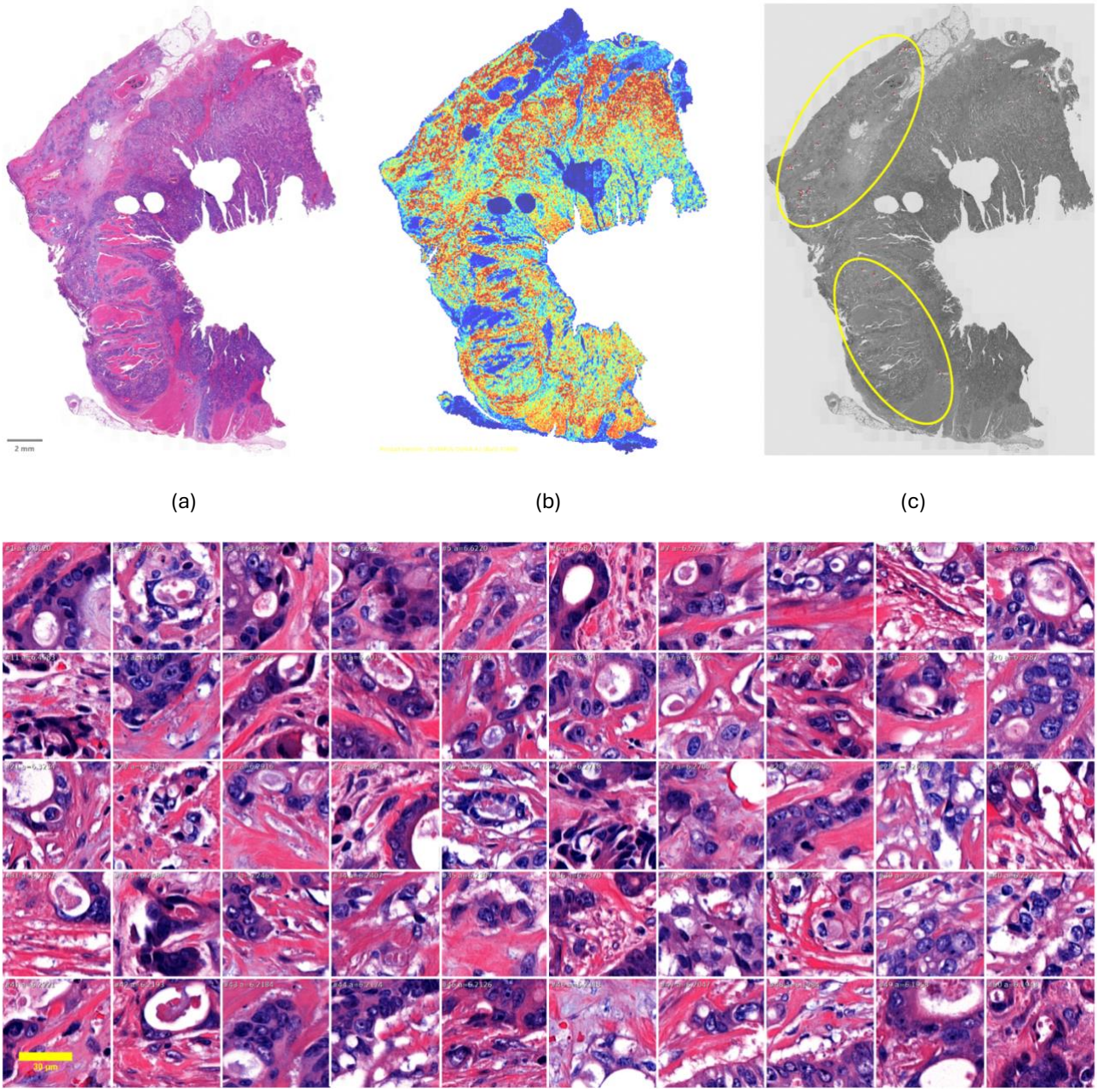


Figure 5.2: Model attention heatmaps highlighting histopathological regions contributing to survival prediction.

Panel (A) shows the H&E-stained tissue section and the corresponding heatmap, highlighting the regions considered most important by the model in dark red and red color. Three boxed areas indicate regions that are further examined in panels (B), (C), and (D). Panel (B) shows a magnified view of rectangle (1) and its corresponding heatmap, illustrating that the dark red and red tiles are mainly located in the tumor and surrounding stroma, while the necrotic area appears in blue. Panel (C) presents a magnified view of rectangle (2) and its corresponding heatmap, where dark red and red tiles are concentrated in the tumor and adjacent stromal regions, whereas the mucin region is highlighted in blue. In panel (D), the magnified view of rectangle (3) and its associated heatmap demonstrates an invasive tumor region in red and dark red, whereas the benign surrounding adipose tissue is dark blue. (Scale bars shown in the lower-left corner of H&E WSI indicates 2 mm; of H&E image in (B) indicates 400 µm; of H&E images in(C), and (D) indicate 200 µm)

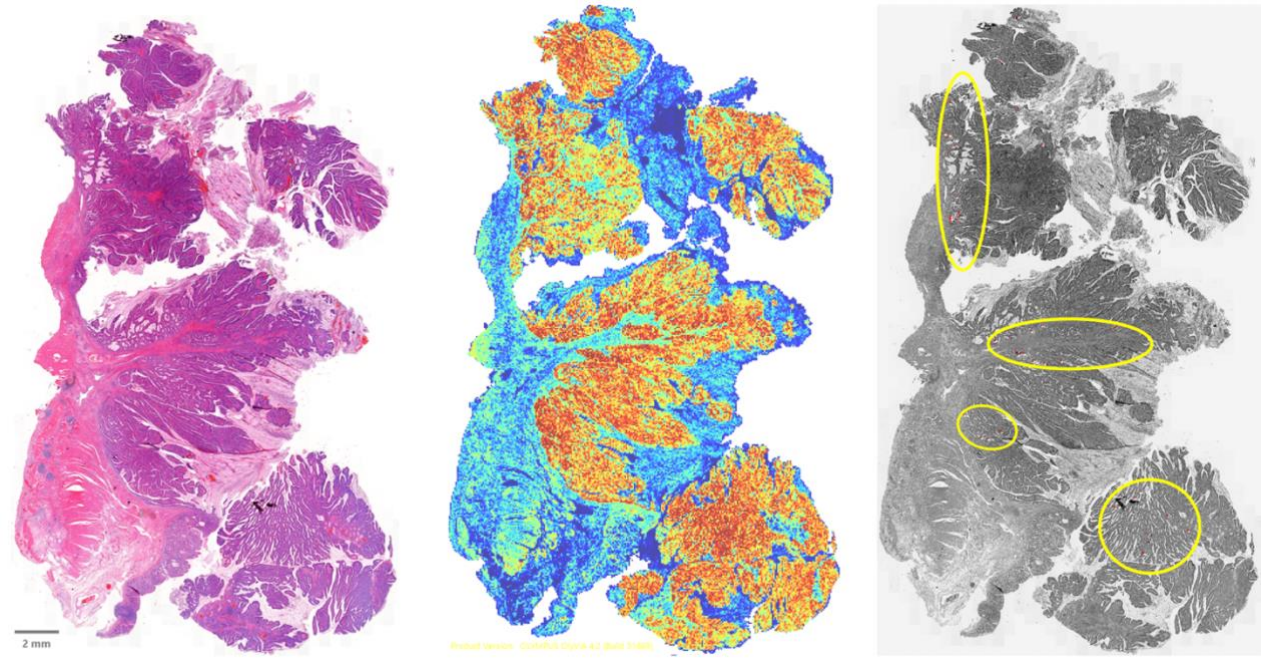


(a) (b) (c)

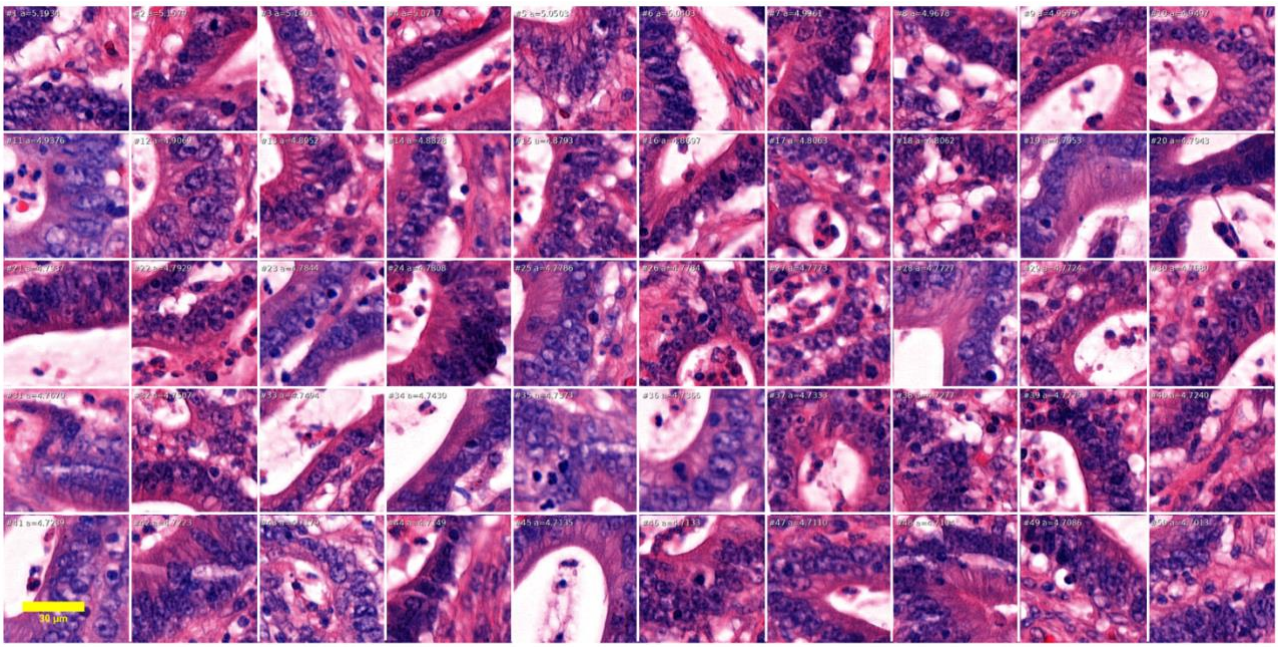
(d)

Figure 5.3: Model attention heatmaps and top-ranked tiles illustrating invasive tumor patterns in a dead-predicted patient.

An H&E tissue section of a patient who died before 5-year endpoint (a). The heatmap demonstrates the high attention areas in dark red, less attention areas in range of color from orange to green, and the non-attention areas in blue (b). The gray scale map shows that the distribution of the top 50 tiles is located in invasion areas (yellow circles) (c). The top-50-tiles grid shows moderate to high-grade nuclei in most of the tiles. The tumor cells form clusters rather than glands, interrupt the keloid stroma (d) (Scale bar shown in the lower-left corner of H&E WSI indicates 2 mm. The grid includes 50 tiles at 40X magnification, tile size 256 x 256 px, scale bar indicates 30 µm).



(a) (b) (c)



(d)

Figure 5.4: Model attention heatmaps and top-ranked tiles illustrating tumor patterns in an alive-predicted patient.

An H&E tissue section of a patient who was alive at 5-year endpoint (a). The heatmap demonstrates that the high attention areas in dark red mostly concentrate in tumor epithelium (b). The gray scale map shows the distribution of the top 50 tiles of a patient who died. Most of the top tiles are located in the tumor area (c). The top-50-tiles grid shows low-grade nuclei in most of the tiles, clear gland structures with intralumen neutrophils (d). (The scale bar shown in the lower-left corner of H&E WSI indicates 2 mm. The grid includes 50 tiles at 40X magnification, tile size 256 x 256 px, scale bar indicates 30 µm).

5.2.2 Top feature clusters by K-means algorithm

All top 10 clusters of each k value (75, 85, 100, 105, 125, and 135) were reviewed and interpreted. Although higher k values yielded increased silhouette scores, the range of silhouette values was narrow (approximately 0.0325 - 0.0399). Therefore, the difference between the highest silhouette value at k = 135 (~0.0399) and the silhouette value at k = 75 (~0.0360) was very slight. The silhouette alone was not the only criterion. The other metrics, including cluster size entropy, normalized entropy, R^2, were evaluated. Although cluster entropy and normalized entropy increased slightly as K grew, the performance of the linear model with forward stepwise selection peaked at K = 75 and decreased for larger K, suggesting that an excessive number of clusters may introduce additional noise rather than providing benefits for prognostic prediction (Table 5.3). Additionally, qualitative examination of the closest tiles in the top 10 clusters suggested that increasing the k value beyond 75 mostly divided histological feature clusters into smaller subclusters and did not improve the ability to reveal new patterns. Therefore, k = 75 was selected as an optimal trade-off between cluster separability, robustness, and interpretability.

Table 5.3: Comparison of cluster size entropy, normalized entropy, and downstream performance (R^2) across different numbers of clusters (k).

K value	75	85	100	105	125	135
Cluster size entropy	4.184	4.342	4.501	4.548	4.730	4.776
Normalized entropy	0.969	0.977	0.977	0.977	0.980	0.974
R^2	0.739	0.734	0.636	0.648	0.484	0.544

After k=75 had been selected, the 15 closest tiles to the centroid of each of the top 10 clusters identified by forward stepwise selection were reviewed and interpreted. The closest tiles were selected from different patients to capture patterns shared across the cohort rather than being dominated by a single patient's characteristics. As the cluster sizes were different, fewer than 15 tiles were available for some clusters. Many clusters were found to correspond to recognizable histological patterns. In particular, several clusters were dominated by stromal areas, including immature or intermediate desmoplastic or fibrous stroma surrounding tumor glands, where the proportion of stroma was greater than that of

tumor cells (Figure 5.5, 5.6). This indicates that the model's high-attention features indeed aligned with biologically meaningful patterns in the tissue.

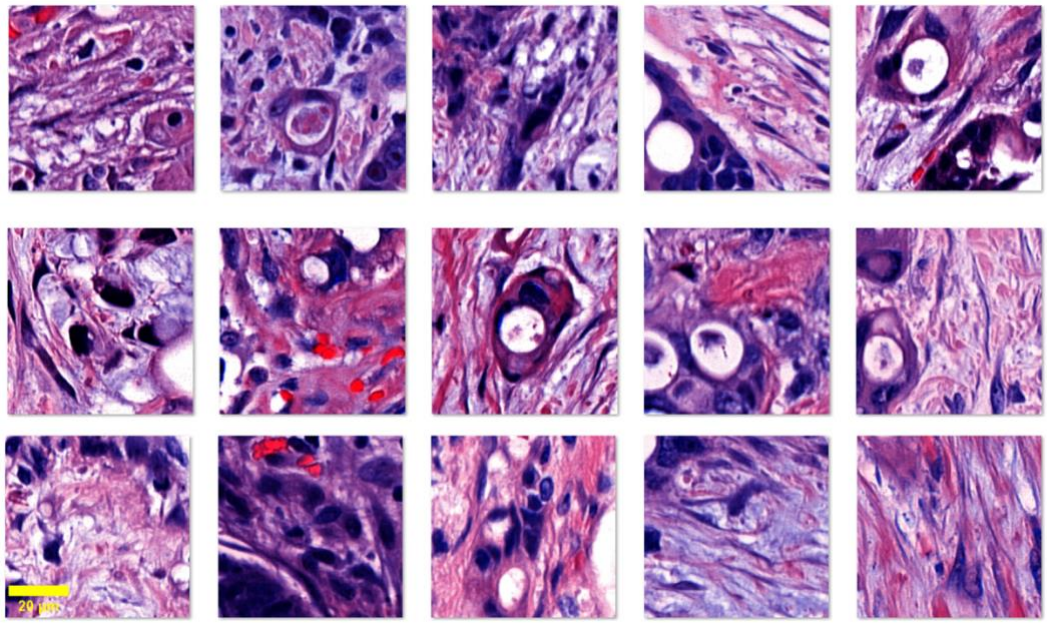


Figure 5.5: Stroma-rich tumor cluster (Cluster #26) with desmoplastic features.

These tiles depict stromal-rich tissue with a stroma ratio ≥50% (stroma-high). The stroma consists of dense fibrous components with elongated, wavy collagen bundles in a pink eosinophilic background, admixed with myxoid material. Spindle-shaped stromal cells with elongated, hyperchromatic nuclei are aligned along collagen fibers, consistent with immature desmoplastic stroma. Scattered atypical round to oval cells are embedded within the stroma, some showing enlarged nuclei and prominent nucleoli. (40× magnification; tile size: 256 × 256 px; scale bar: 20 µm)

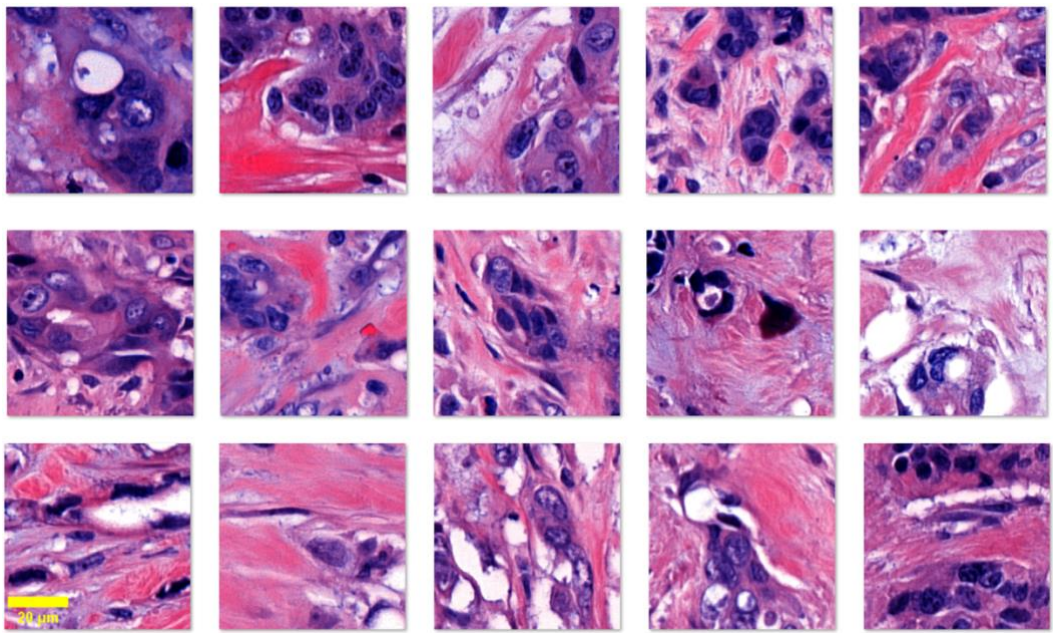


Figure 5.6: Stromal-rich tumor cluster (Cluster #15) with keloid-like desmoplasia and nuclear atypia

These tiles represent higher cellular, keloid desmoplastic tumor regions with marked nuclear atypia and the stroma ratio ≥ 50% (stroma-high), reflecting aggressive histomorphological features. (40X magnification, tile size 256x256 px, scale bar indicates 20 µm).

Few clusters were collections of pale color patches or purely collagen bundles (Figure 5.7), lacking a clear, meaningful morphological identity, which were acknowledged as a limitation in our post-hoc analysis.

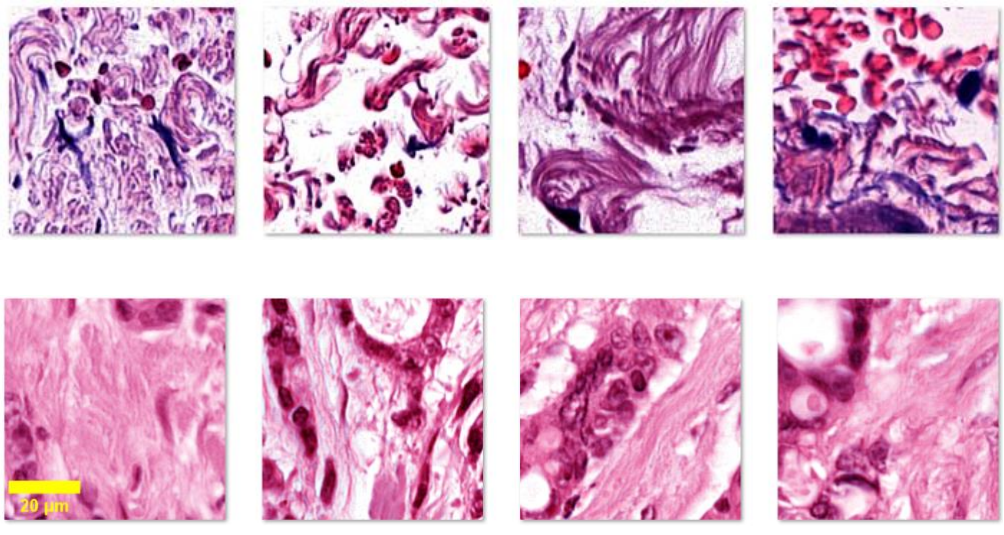
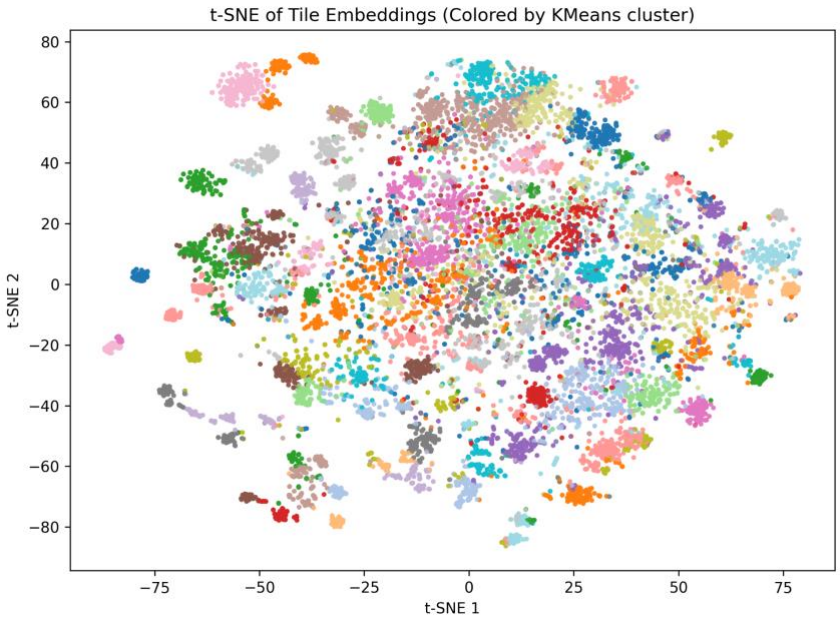


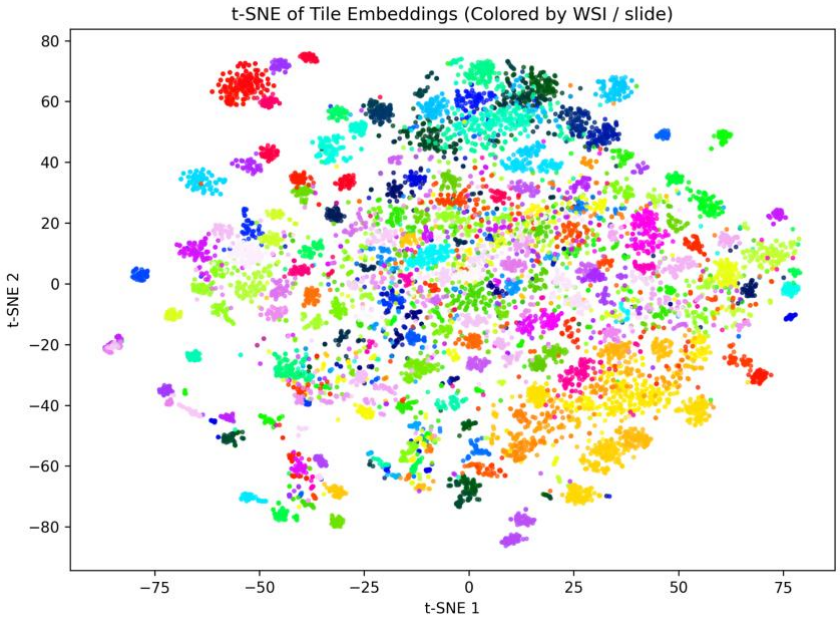
Figure 5.7: Representative cluster tiles highlighting collagen features (a) and residual staining variability not fully corrected by color normalization (b). (40X magnification, tile size 256x256 px, scale bar indicates 20 µm).

5.2.3 Dimensionality reduction and visualization

When the tile embeddings were colored by K-means clusters, tiles from the same cluster formed compact groups in t-SNE space, showing that the clustering captured similar feature representations (Figure 5.8a). In contrast, when the same embeddings were colored by slide ID, tiles from the same slide were spread across multiple regions and did not form clear slide-specific groups (Figure 5.8b). This suggests that the learned features were driven mainly by morphology rather than slide identity or technical variation, and that the observation was consistent across both projection methods.



(a)



(b)

Figure 5.8: t-SNE visualization of K-means clustering and WSI distribution (k = 75)

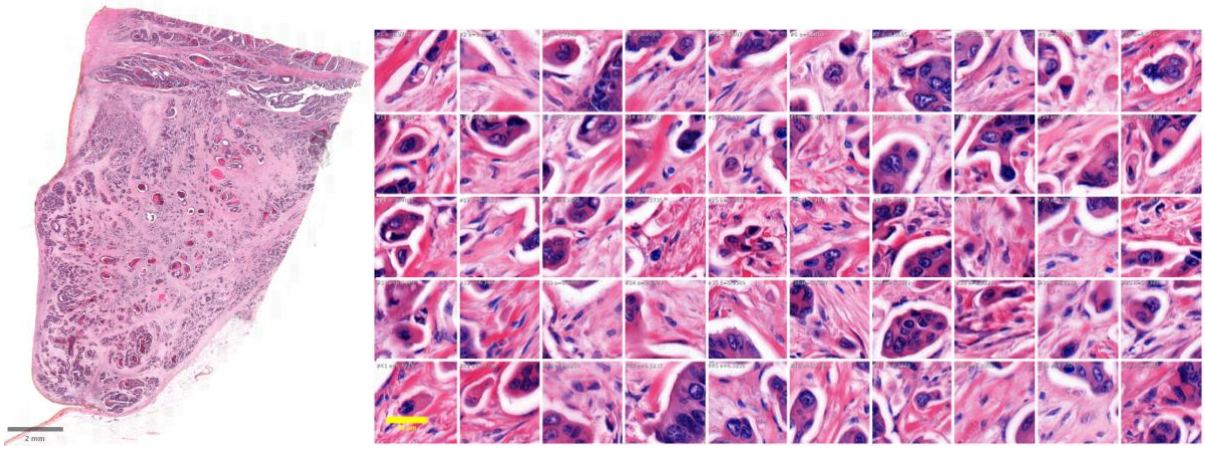
t-SNE visualization colored by K-means clustering (a) and by WSIs (b) with k = 75. The results show that slides with similar features are grouped into the same clusters by the K-means algorithm.

5.3 Error Analysis

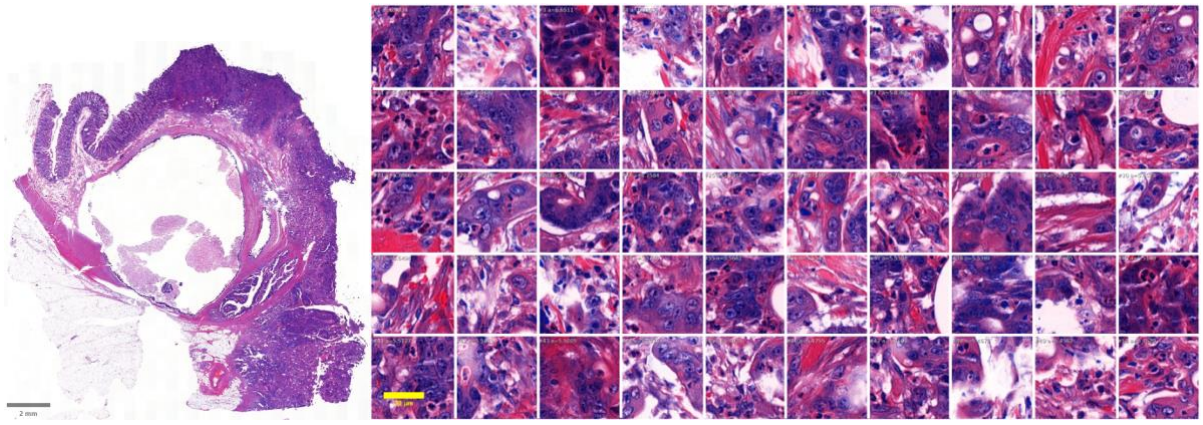
5.3.1 Histopathological patterns in misclassified cases

Among misclassified cases, the common patterns in top tiles in false dead-predicted cases for 5-year survivors were quite similar to the true dead cases. They showed a stroma-high feature, small clusters of tumor cells rather than typical gland structures (Figure 5.9A).

Conversely, in false alive-predicted cases in which patients died within 5 years, top tiles showed a stroma-low feature, and a high prevalence of neutrophil-infiltrate tiles (Figure 5.9B).



(A)



(B)

Figure 5.9: Representative top-ranked tiles from misclassified cases illustrating variations in stromal composition and tumor architecture.

Panel (A) includes WSI and a grid of the top 50 tiles of a patient who lived more than 10 years after surgery, but the prediction was dead at the 5-year endpoint. The top tiles show a high ratio of stroma, though the stroma is less aggressive than the true dead cases (no myxoid, few keloid collagen bundles), and small tumor cell clusters. Panel (B) presents a tissue section and its top-tile grid of a patient who was predicted as alive at the 5-year endpoint by the model, though the patient died 3 years after the operation. The top 50 tiles show tumor cell clusters or small gland structures surrounded by stroma and inflammatory cells. The proportion of stroma is lower than in dead-prediction cases. (Scale bars shown in the lower-left corner of H&E WSI indicate 2 mm. Each grid includes 50 tiles at 40X magnification, tile size 256 x 256 px, scale bars indicate 30 μm.)

5.3.2 Intra-patient heterogeneity and aggregation effects

As described in the Methods section, the tissue segmentation and slide splitting step separated disconnected tissue fragments within the WSI into independent image regions if no physical continuity was detected. Consequently, a single patient could be represented by

multiple segmented tissue pieces, each of which was processed separately by the model and assigned an individual predicted probability.

In several misclassified cases, these separated tissue regions from the same patient yielded slightly different predicted probabilities. Because a classification threshold of 0.5 was applied at the image level, some tissue pieces were classified as dead-predicted (probability > 0.5), while others were classified as alive-predicted (probability < 0.5). This discrepancy was particularly evident in borderline cases, where subtle morphological differences between tissue fragments led to probability estimates close to the decision boundary.

When aggregating predictions to the patient level, the final probability was computed using the predefined mathematical aggregation function. However, when individual image-level probabilities were distributed around the threshold, the aggregation process often resulted in a moderate or attenuated final probability. While this continuous probability may still contain meaningful risk information, the subsequent conversion into a binary outcome (alive vs. dead) potentially amplified small numerical differences and contributed to misclassification. Example in Figure 5.10 illustrates this effect.

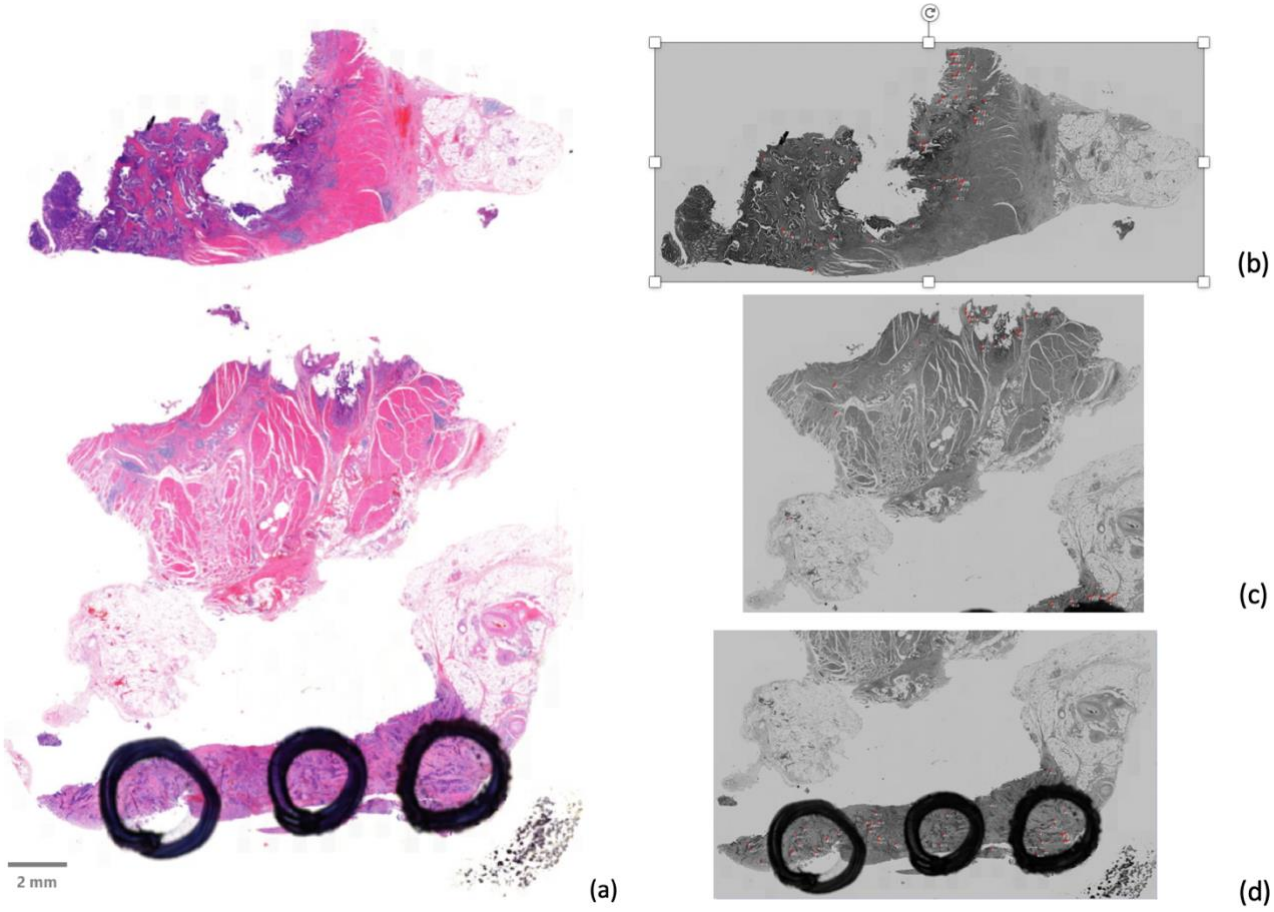


Figure 5.10: Image-level predictions illustrating discordant outcomes contributing to patient-level misclassification.

The WSI of one patient who was surviving at the 5-year end point includes three tissue sections, which were separated into three images in the preprocessing steps (a). The first and second sections were predicted as alive with probabilities of 0.5400 and 0.5785, respectively (b, c). The third section had a death probability of 0.7108, which outweighed the alive probability of other sections, resulting in the dead prediction at the patient-level outcome (d) (Scale bar shown in the lower-left corner of H&E WSI indicates 2 mm)

5.3.3 Attention to irrelevant or artifactual regions

While the tissue segmentation algorithm succeeded in removing the blank background in most cases, there were still some images including blank space inside the tissue section. Among those cases, occasionally, some blank tiles had quite high attention scores. As a result, these meaningless tiles contributed to the predicted probability, leading to a wrong prediction, as was shown in Figure 5.11.

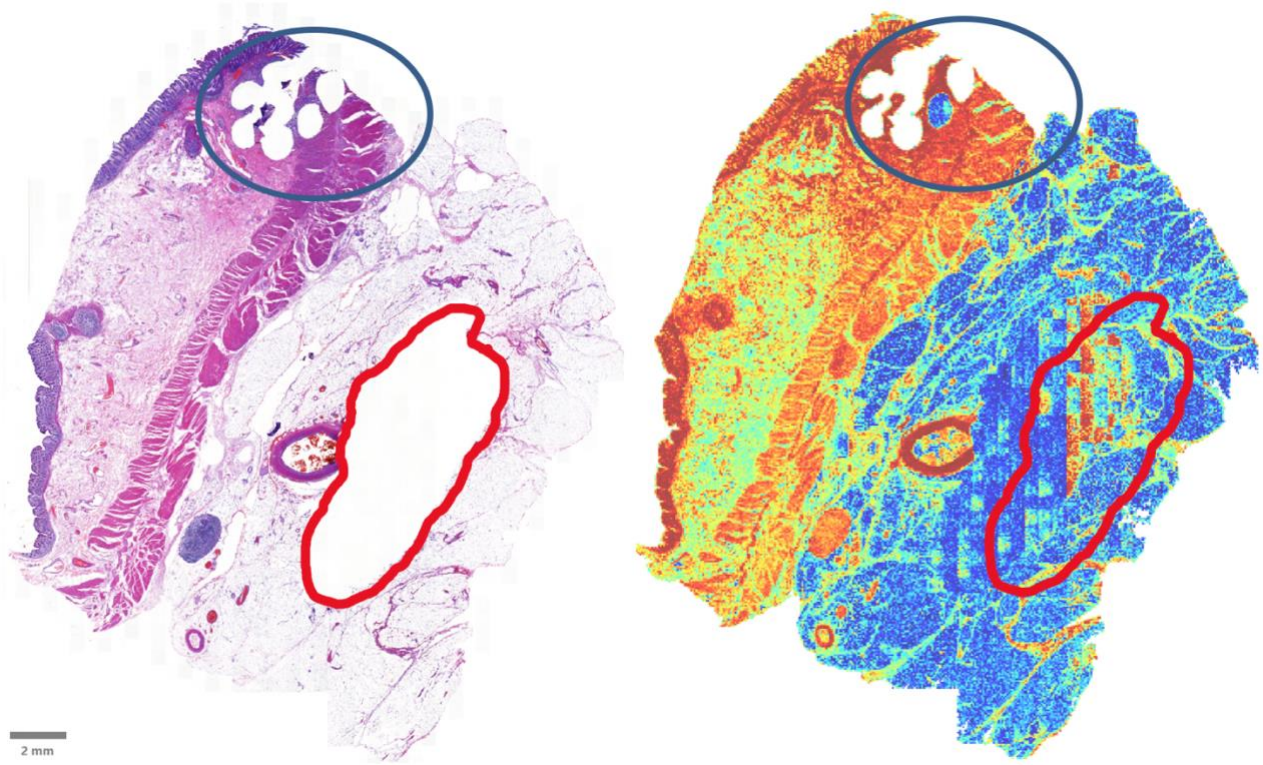


Figure 5.11: Attention heatmap illustrating background artifacts and incomplete tissue exclusion affecting model predictions.

The red circles in the heatmap correspond to the total blank area in the WSI (left). However, in the heatmap, there are a few tiles in red, yellow, or green colors instead of blue, as they should be. The top background area (blue circle) was correctly excluded, but a round white region was still retained in the heatmap and appeared in blue (Patient was predicted alive, though had died at the 5-year time point) (Scale bar shown in the lower-left corner of H&E WSI indicates 2mm).

5.3.4 Data-related limitations

Missing informative spots due to TMA cores

The dataset was collected over 27 years, several tissue blocks had been used for other studies using tissue microarray (TMA) (Figure 5.12). In routine practice, the TMA cores are the areas that pathologists consider most important for the aim of studies. Besides, some tissue sections were partially missing due to tissue preparation artifacts (Figure 5.14). This study was recently conducted on WSIs, so only the remaining tissue sections could be analyzed. It was understood that this issue might have affected model performance because some highly informative regions were missing.

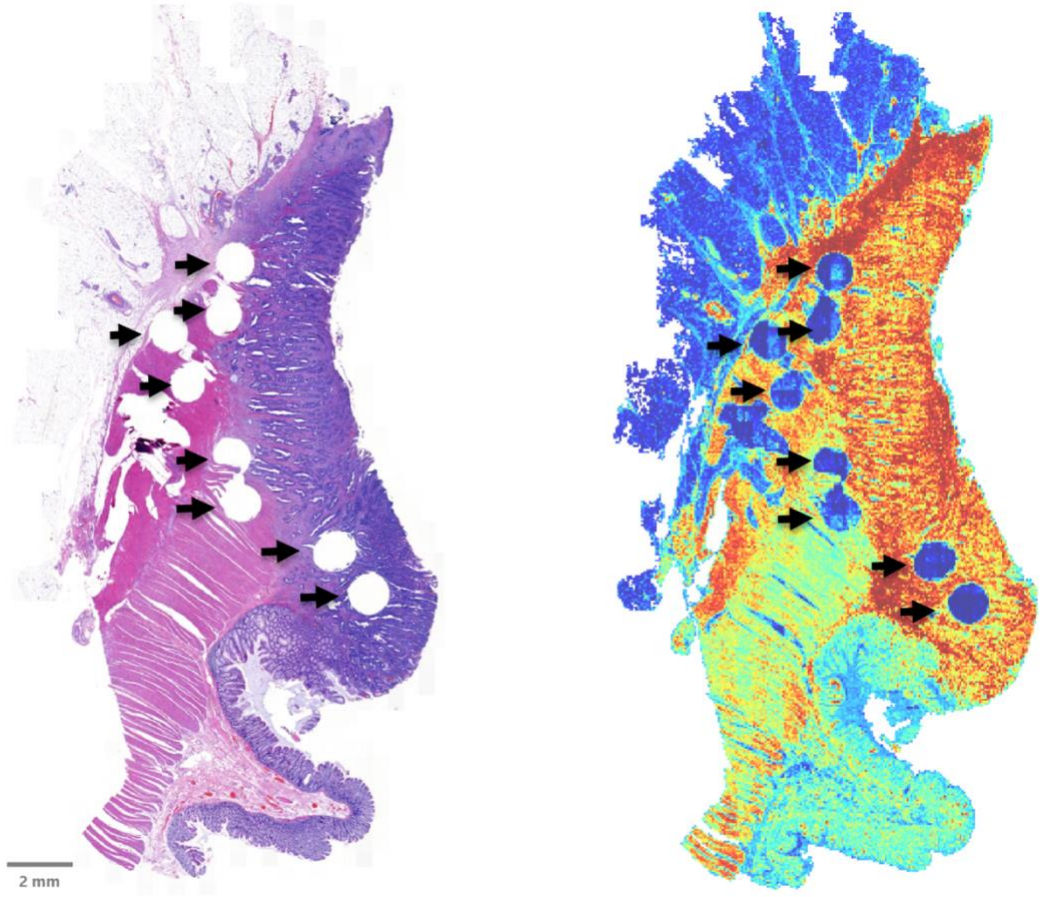
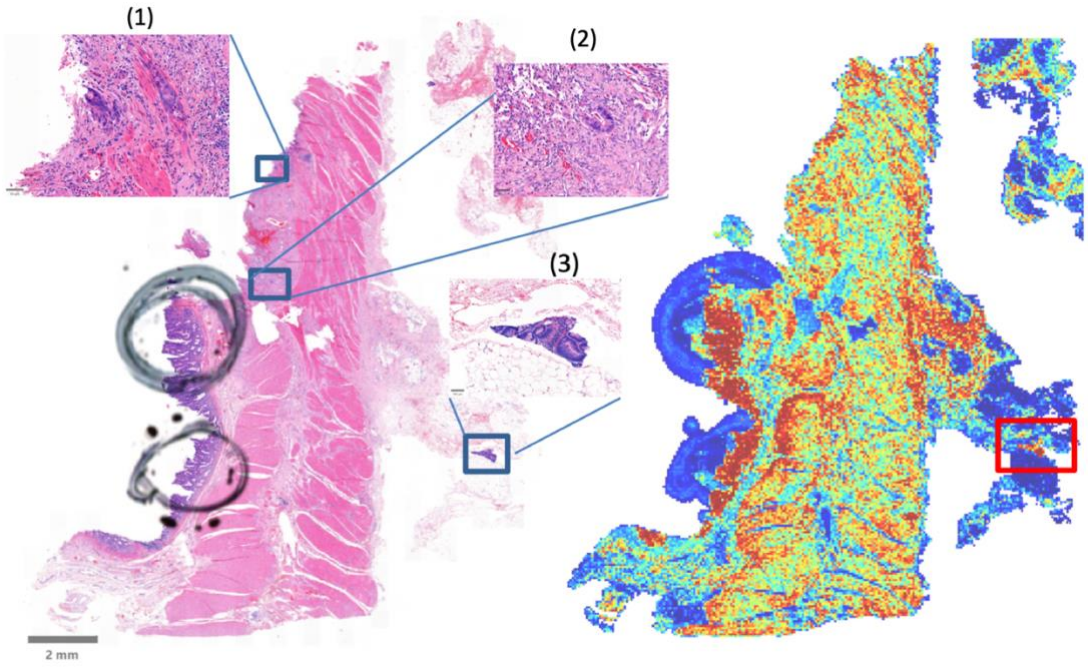


Figure 5.12: Attention heatmap illustrating missing tissue regions within high-attention areas and their potential impact on model predictions.

The WSI and corresponding heatmap show a tissue section containing several round missing spots (black arrows). In the heatmap, these blank spots are in blue, which means that the model can realize the non-informative regions for prediction. However, most of these spots are located within the high-attention area (dark red, red or orange), suggesting that potentially informative tissue was removed and could have contributed to the final prediction. (patient was predicted alive, though had died at the 5-year time point) (Scale bar shown in the lower-left corner of H&E WSI indicates 2 mm).

Non-representative slide selection

The digital slide images chosen in the dataset varied. For most patients, only one slide was selected for model training (a small number had two to three slides), although in clinical practice, each typical case contains several tissue sections capturing different tumor regions, surgical margins, and lymph nodes when available. Therefore, the dataset contains a variety of histological types across the patients: some cases included only lymph node sections with metastatic tumor involvement (Figure 5.14), while some others consisted of the peripheral sections or only the polyp, the tumor epithelium without the deeper layers (Figure 5.13). These images might provide very aggressive information (invasion, destroying lymph nodes structure) or very mild pattern (tumor was limited in muscularis propria). This variability may affect the model learning and performance.



(A) (B)

Figure 5.13: Attention heatmap and H&E image illustrating mild tumor morphology and spurious model attention to preparation artifacts in a misclassified case.

(A) H&E WSI of a patient, who died 4 years after surgery but was predicted as alive at 5-year end point, shows a mild morphology where most of the tumor is limited in epithelium, only tiny invasion areas (1,2) with mature desmoplastic reaction in the muscularis propria. There is a small cluster of tumor glands that appear because of preparation artifacts (3). Overall, this WSI does not show aggressive features clearly. (B) The association attention-based heatmap on the right shows that the model excluded the ink marks and blank, benign fat tissue in prediction. The red rectangle depicts the preparation artifact where the model puts attention weight on the displaced tumor glands as shown in (3) rectangle in the WSI. (Scale bar shown in the lower-left corner of H&E WSI indicates 2 mm).

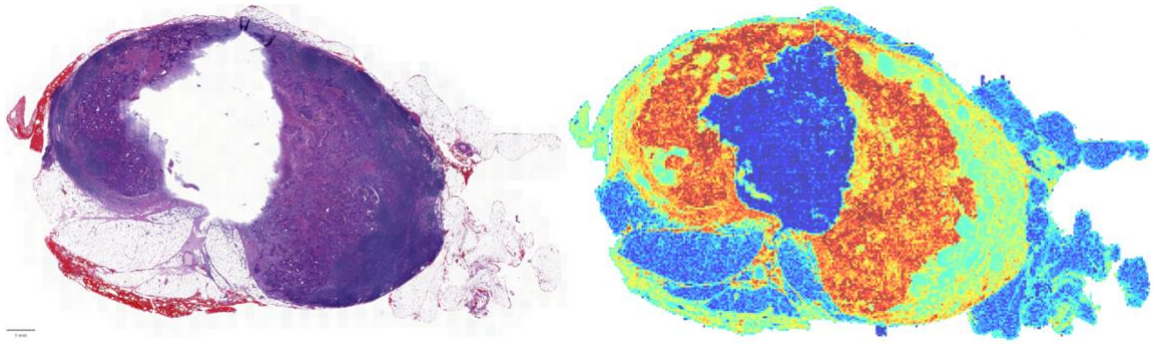


Figure 5.14: H&E image and attention heatmap of a metastatic lymph node highlighting limited primary tumor information and missing tissue regions contributing to misclassification.

This slide could not provide the information of the original tumor as well as the invasion, desmoplastic stroma. The WSI also shows partial missing areas in the centre of the lymph node, which is in dark blue in the heatmap. (Patient was predicted alive though had died at 5-year time point) (Scale bar shown in the lower-left corner of H&E WSI indicates 1 mm).

6 Discussion

6.1 Model performance

6.1.1 Model evaluation

The UNI2-h–CLAM survival model achieved robust discrimination of 5-year outcomes in CRC, with performance that is comparable to recent foundation, MIL, morphology-aware models. Its performance was also clearly stronger than earlier TMA-based work on the same Helsinki cohort [11]

Using UNI2-h features with CLAM and Macenko normalization, the model achieved an image-level AUC of approximately 0.75 (±0.02), accuracy ~0.70, F1-score ~0.57, sensitivity ~0.55, and specificity ~0.79. Applying stain color normalization (Macenko's method) yielded only marginal improvements in performance (image-level AUC increased from ~0.74 to 0.75, accuracy 0.68 to 0.70), suggesting that the model retained some dependence on residual stain variation. This is plausible, as the large number of slides required staining to be performed in multiple batches, and variations between staining batches, together with manual processing, likely introduced substantial variability in color appearance [55]. Unsupervised K-means clustering of tile features still produced several clusters defined primarily by color tone, indicating that color differences across slides were only partially corrected. Future work should investigate more advanced stain normalization techniques or tuning the Macenko algorithm parameters to further reduce color-driven bias [68, 69].

Patient-level aggregation substantially improved performance. By averaging predicted probabilities across multiple slides per patient, the patient-level AUC increased to 0.79, with accuracy 0.73, F1-score 0.62, sensitivity 0.62, and specificity 0.80. This improvement indicates that aggregating multiple sections per patient increases robustness. This aggregating step mirrored how pathologists combine findings from all available slides to reach a holistic assessment. The improved sensitivity suggests that slide aggregation helped capture more true positive cases while maintaining high specificity.

The concordance index (C-index) for the CLAM risk scores was 0.738, indicating good agreement between the model's predicted risk ranking and the actual survival times. Furthermore, Kaplan–Meier survival stratification based on the model's prediction was highly significant: patients who were predicted as "alive" showed markedly longer actual survival than those predicted as "dead", with well-separated survival curves ($p < 0.001$) and an HR of 4.27 (95% CI: 3.22–5.66). This clear stratification confirms that the model's predictions correspond to clinically meaningful outcomes. The use of both threshold-based metrics (AUC, accuracy, F1, sensitivity, specificity) and rank-based discrimination (C-index) aligns with evaluation strategies in recent studies, facilitating cross-study comparison.

6.1.2 Comparison to earlier TMA-based work on the same cohort

Bychkov et al. (2018) trained a CNN+LSTM model on H&E-stained TMAs from 420 CRC patients at HUS, achieving an AUC of 0.69 and an HR of 2.3. In contrast, this study leveraged 1,381 patients and 1,407 whole slide images (1,950 tissue images) from the same institution, moving from sparse TMA sampling to near-whole-tumour representation. The higher AUC (0.79 vs 0.69) and larger HR (4.27 vs 2.3) suggest that combining foundation-model features with attention-based MIL on WSIs captures more prognostic information than earlier TMA-based deep learning approaches [11].

6.1.3 Influence of dataset scale and composition

Compared with the PRISM (Sajjad et al.,2025), which was trained on 424 stage III CRC patients only, the UNI2-h–CLAM model covered a broader clinical spectrum of 1,381 patients across stages I–IV. PRISM reported an AUC of 0.70 ± 0.04 and an HR of 3.34 for 5-year overall survival (OS), which is slightly lower than the values observed here. However, direct comparison requires caution, as PRISM used all-cause mortality (overall survival) for stage III patients, while the present model focused on death-related CRC within 5 years after operation across all stages. This restricted endpoint avoided the noise introduced by deaths from unrelated causes, which histology cannot predict. This cleaner outcome signal may partly account for the superior metrics observed here [16].

The Foundation model of Wang et al. (2025) pooled 2,594 CRC patients across multiple international cohorts and achieved C-indices of roughly 0.714 – 0.757 for DSS in the

colorectal dataset [14]. Jiang et al. (2024) reported a C-index of 0.78 (0.70 – 0.84) for DSS in a multicenter study of 4,428 patients [4]. These results are comparable to the C-index of 0.738 achieved here, despite the present model being trained on a single institutional cohort.

6.1.4 Workflow strategies

Several recently published studies shared similar strategies combining foundation models and MIL. Li et al. (2022) introduced a distribution-based multiple instance learning approach (DeepDisMISL), a distribution-based MIL approach using Xception features, achieving a C-index of 0.638 (0.626 – 0.66) on large international cohorts [46]. Jiang et al. (2024) used RetCCL, a self-supervised histology-specific encoder, with attention MIL for patient-level survival prediction [4]. Wang et al. (2025) fine-tuned a foundation model pretrained on over 100,000 WSIs in a weakly supervised framework [14]. PRISM, in turn, combines morphology-informed features from expert-annotated tissue regions with foundation model embeddings and MIL aggregation. This is an approach closely related to the present study, except for the explicit morphology-informed classifier [16].

Compared to these approaches, the UNI2-h–CLAM model demonstrates that integrating readily available foundation-model features with a relatively simple attention-based MIL head and standard preprocessing is sufficient to achieve competitive performance for 5-year survival prediction in a single-centre CRC cohort, without task-specific pretraining, extensive annotation, or multi-centre data.

6.2 Interpreting the deep learning model through post-hoc morphological analysis

6.2.1 Post-hoc interpretation methods

The most challenging aspect in applying deep learning to computational pathology is the lack of interpretation. While they can achieve strong prognostic performance, understanding which histological features drive predictions remains difficult. Bychkov et al. (2018) used the same Helsinki cohort but offered no interpretation of the features underlying their predictions [11]. This thesis builds upon the earlier work by demonstrating that, through foundation model embeddings (UNI2-h), attention-based MIL, and systematic clustering, it is possible not only to predict outcomes but also to provide interpretable explanations of the model's prognostic rationale.

In this thesis, we addressed this issue through a multi-step post-hoc interpretation pipeline that combined attention heatmaps, foundation model (UNI2-h) feature embeddings, K-means clustering, and qualitative pathologist review. This approach was directly inspired by Wulczyn et al. (2021), who used clustering-derived histologic features to explain the variance in deep learning system (DLS) predictions for CRC survival. Wulczyn et al. clustered tile embeddings into 200 clusters and demonstrated that standard clinicopathologic features (T-category, N-category, grade) explained only 18% of the variance in DLS scores, whereas 200 clustering-derived features explained 73–80%, and a subset of 10 features selected by forward stepwise regression captured 57–61% [10].

This study followed a conceptually similar strategy but differed in several respects. We used UNI2-h, a pathology-specific foundation model, rather than an image-similarity model [70] for embedding extraction. Attention-based MIL was applied to select the most informative tiles, and silhouette analysis combined with downstream regression metrics was employed to determine the optimal number of clusters. At $k = 75$, our top 10 clusters achieved an R^2 of 0.739, indicating that these morphological features captured a substantial portion of the model's decision logic. This value exceeds Wulczyn et al.'s 0.57–0.61 for their top 10 features, although the comparison must be interpreted cautiously given differences in dataset size, DL model architecture, and embedding models [10].

The clustering-based approach used here has the advantage of not presupposing which cellular or microenvironmental features are important. Prognostic clusters emerged purely from UNI2-h embeddings and their association with survival, and only afterward are they interpreted in terms of desmoplastic reaction (DR), tumor differentiation, and tumor-stroma ratio (TSR).

Zhou et al. (2023) combined WSIs, clinical variables, and mutation signatures and used heatmaps reviewed by pathologists to identify morphological features associated with risk. They found that nuclear size pleomorphism, intense cellularity, and abnormal structures were related to high-risk cases, while regular and smaller cells were linked to low-risk cases. While the first part of Zhou et al.'s strategy is similar to the approach used here, their method explicitly assumed that cell counts, cell types, and nuclear size are key descriptors by designing a HoverNet segmentation model. The present clustering-based approach, in

contrast, identifies prognostically relevant patterns without predefined features. However, running a nucleus segmentation model over the high-attention clusters, as Zhou et al. did, would be a natural extension of the current framework, transforming the qualitative cluster interpretation into a quantitative one [41].

Wei et al. (2024) used Grad-CAM to generate heatmaps and grouped high- and low-risk regions into four categories. Pathologist inspection identified immature tumour stroma, disorganised gland structures, and small clusters of poorly differentiated cells as high-risk features, while infiltrating inflammatory cells were linked to better prognosis. Furthermore, Wei et al. (2024) linked Colorectal Cancer Risk Score (CRCRS) to multiomics thus extending interpretability from morphology into underlying tumour biology. Our unsupervised, embeddingspace clustering is more datadriven and systematic in how it defines prognostic morphologies, whereas Wei et al. provided a more direct and biologically integrated interpretation [9]. Therefore, the two approaches are complementary rather than one simply superseding the other.

Jiang et al. (2024) focused primarily on demonstrating robust prognostic performance across four cohorts, with attention heatmaps and top tiles shown only for six selected high-risk cases. Their results showed that the model primarily focused on poorly differentiated tumor regions and areas of fat tissue invaded by the tumor [4]. Wang et al. (2025) relied mainly on transformer-based attention heatmap at the slide level, showing that high-risk areas contain more tumor cells and fewer immune cells [14]. Our study complements their work by providing systematic morphological interpretation, showing which tissue patterns (desmoplastic stroma, tumor differentiation, and inflammatory infiltration) deep learning models used for survial prediction.

6.2.2 Features associated with poor prognosis

The attention heatmaps revealed that the model concentrated on tumor epithelium, invasion regions, and peritumoral stroma, while largely ignoring necrosis, mucinous areas, and benign tissue. Among dead-predicted cases with a high level of confidence, the most informative tiles consistently showed moderate-to-high-grade nuclear atypia, poorly differentiated or solid tumor growth, and marked immature (myxoid) or intermediate (keloid-like)

desmoplastic stroma, with a qualitative stroma ratio exceeding 50% in the majority of cases. Spatially, high-attention tiles in dead-predicted cases were concentrated in invasion zones.

These features align closely with established poor-prognosis histological factors in CRC. Dead-predicted tiles often show moderate to high-grade nuclear atypia and poorly differentiated or solid growth, reflecting a high histological grade. The World Health Organization (WHO) 2019 and AJCC 8th guidelines classify these as adverse features, especially in stage II disease, where poor differentiation is a major high-risk criterion [19]. Zhou et al. (2023) similarly found that high-risk tiles exhibited nuclear size pleomorphism, intense cellularity, and abnormal structures, with quantitative analysis showing significantly higher cellularity and larger tumor nuclei compared to low-risk tiles [41].

The model concentrated on small clusters of tumor cells with poorly formed glands at the invasive edge and in pericolonic fat. These features overlap morphologically with the “poorly differentiated clusters” and tumor budding at the invasive front, known to predict recurrence and poor survival [19, 23, 71]. Similarly, Wei et al. (2024) independently identified immature tumor stroma, poorly formed gland structures, and small tumor cell clusters as unfavorable features in their CRCRS using Grad-CAM visualization [9].

Most notably, the model's emphasis on immature and intermediate desmoplastic stroma with a high stromal fraction mirrors emerging evidence that DR type and TSR are independent prognostic markers [72, 73]. Akimoto et al. (2022) demonstrated that patients with mature desmoplastic reaction (DR) had more favorable CRC-specific survival outcomes than those with immature DR [72]. Similarly, Hu et al. (2023) reported that mature DR is associated with the best 5-year OS rates (85%), while the survival rates decreased to 75.3% and 65.2% for intermediate and immature reactions, respectively [73]. These patterns may also overlap with the consensus molecular subtypes 4 (CMS) of CRC, which is characterized by prominent stromal infiltration and the worst survival [74]. The desmoplastic stroma identified here thus potentially reflects an underlying molecular phenotype without requiring genomic analysis.

The model's attention to stroma-rich regions is also consistent with the prognostic significance of TSR. Abbet et al. demonstrated that automated TSR was a statistically significant predictor of overall survival in CRC [47]. Although TSR was never supervised in the present model, the model still consistently focused on stroma-rich regions in poor-outcome

cases, suggesting that TSR-related information was learned as part of the survival prediction task.

6.2.3 Features associated with favorable prognosis

In contrast, survival-associated tiles predominantly displayed preserved glandular architecture, low-to-moderate-grade nuclei, and frequent inflammatory cell infiltration, consistent with better-differentiated tumors and favorable outcomes according to WHO classification or College of American Pathologists (CAP) criteria [19]. In contrast to most recent studies and current criteria that highlight lymphocyte infiltration as a favorable prognostic marker, the top tiles in this study were dominated by acute inflammatory cells, including neutrophils and eosinophils [19]. Wei et al. (2024) also identified infiltrating inflammatory cells as the key favorable prognostic feature in CRCRS [9], while Zhou et al. (2023) found that low-risk regions showed more regular and smaller cells [41].

6.2.4 Clustering validation through dimensionality reduction

The t-SNE analysis confirmed that the clustering of tile embeddings was driven by morphological similarity rather than slide-level characteristics. When colored by K-means cluster assignment, tiles formed coherent groups in both projections; when colored by slide identity, tiles from the same slide were widely distributed without forming slide-specific clusters. This demonstrates that the UNI2-h foundation model learned generalizable, morphology-based features shared across WSIs rather than memorizing staining variations or technical artifacts, strengthening confidence in the biological relevance of the identified clusters.

6.3 Error analysis

6.3.1 Morphological ambiguity in misclassified cases

False dead predictions among 5-year survivors displayed morphological patterns remarkably similar to true dead cases, with a high TSR and tumor cell clusters without typical gland structure. Conversely, false alive predictions in patients who died showed a lower TSR and a high prevalence of neutrophil-infiltrate tiles. Zhou et al. (2023) met a similar issue with their model when further investigating misclassified cases. The false “high-risk” cases had higher

cell densities and a higher number of cancer cells [41] .The false survival cases suggest that the model learned a simplified association between inflammatory infiltration and favorable outcomes, without distinguishing between prognostically favorable lymphocytic infiltration and other inflammatory cell populations. Akimoto et al. (2022) showed that immune cell type determines prognostic significance, with memory cytotoxic T cells and M1-like macrophages conferring survival benefit, while other inflammatory cell populations may not [72]. The model's inability to make this distinction represents a fundamental limitation of the current tile-level analysis.

These misclassification patterns reflect genuine morphological overlap between prognostic categories rather than random errors. Song et al. (2020) found that deep learning models and pathologists made similar types of diagnostic mistakes, suggesting a shared "reasoning logic" when facing ambiguous cases. The cases where stromal features mimicked aggressive morphology in long-survivors, or where neutrophilic inflammation masked poor prognosis, represent the same diagnostic ambiguities that challenge human pathologists. This indicates that these apparent misclassifications may reflect genuine biological variability in tumor behavior, influenced by additional host and microenvironmental factors beyond what can be captured by histology alone, rather than representing simple model failure [75].

6.3.2 Intra-patient heterogeneity and technical artifacts

The tissue segmentation procedure occasionally separated disconnected tissue fragments into independent regions, yielding discordant predictions from the same patient. Binary thresholding at 0.5 then amplified small numerical differences near the decision boundary, contributing to misclassification. Additionally, blank spaces within tissue sections sometimes received inappropriately high attention scores, contaminating predictions. Song et al. (2020) similarly noted artifact sensitivity in their model and emphasized the importance of preprocessing quality control [75]. This finding reinforces the need for robust artifact detection beyond simple background removal.

6.3.3 Data-related limitations

The dataset was collected over 27 years, during which tissue blocks had been used for prior TMA studies, creating missing informative regions. Slide selection variability meant that some

patients were represented by metastatic lymph node sections while others had only peripheral tumor sections, introducing heterogeneity that affects both model learning and interpretability. These limitations are particularly notable given our single-centre cohort of moderate size, compared to the multicentre datasets of Wulczyn et al. (3,652 cases), or Jiang et al. (4,428 patients), which provided more robust validation [4, 10].

We also acknowledged that a few top clusters corresponded to pale patches or purely collagen bundles without clear morphological identity. Wulczyn et al. similarly noted that approximately 20% of DLS variance remained unexplained by their 200 clusters, suggesting that some prognostically relevant features may not be readily captured by unsupervised clustering of image embeddings or may represent subtle features below the threshold of qualitative pathologist interpretation [10].

7 Limitations and future directions

7.1 Limitations

This study has some limitations that should be considered when interpreting the findings. The model was developed and validated on a single-centre cohort, collected over a period of 27 years (1982–2009). This long collection period likely introduced variability in staining protocols and tissue preparation that could not be fully resolved by Macenko's color normalization. Additionally, tissue blocks previously used for TMA studies resulted in partial tissue loss on some slides, and inconsistent slide selection across patients, which may have affected both model learning and interpretability

In this study, we excluded non-related CRC deaths within five years rather than retaining them as censored observations, which deviates from the conventional handling of non–cancer-specific deaths in DSS analysis. This choice simplified the outcome definition for model training but reduced the effective sample size and may introduce selection bias if these patients differ systematically from the remaining cohort.

Finally, the post-hoc morphological interpretation relied partly on qualitative visual assessment, and a small subset of clusters could not be assigned a clear histopathological identity. In addition, the tile-level representation did not allow reliable differentiation between immune cell subtypes, which constrains biological interpretability given the known prognostic importance of immune cell composition.

7.2 Future directions

Several avenues could further strengthen and extend this work. First, future studies should adopt a conventional DSS framework in which deaths from non-CRC causes are retained as censored observations rather than excluded and validate whether the model maintains its performance under this more standard survival definition. Second, more advanced stain normalization strategies, together with robust artifact detection and standardized slide selection, could reduce color and technical bias in cohorts collected over long time periods. Third, integrating automated TSR and DR scoring, as well as cell-level segmentation, would allow the qualitative cluster findings to be quantified and compared directly with established

histopathological scoring systems. Fourth, linking the model-derived morphological patterns to genomic and transcriptomic data could clarify whether stromadominated clusters correspond to CMS4 and other molecular subtypes, and help explain why certain patterns confer poor prognosis. Finally, external validation in large, multicentre CRC cohorts is needed to test the generalizability of the identified stromal and inflammatory signatures across different laboratories, scanners, and patient populations.

8 Conclusion

The results demonstrate that the UNI2-h–CLAM survival model is capable of stratifying colorectal cancer patients into distinct prognostic groups based on H&E-stained whole slide images.

Model interpretation further suggests that the learned features correspond to biologically meaningful histopathological patterns, particularly tumour architecture, invasion-related regions, and stromal responses, which are consistent with established prognostic factors in colorectal cancer.

At the same time, several limitations were identified, including sensitivity to tissue heterogeneity, non-representative sampling, and occasional attention to artefactual or non-informative regions, all of which may affect prediction reliability.

In conclusion, this work contributes to the evidence that unsupervised deep learning survival models for CRC are concordant with known prognostic markers. By providing a systematic, data-driven interpretation framework that goes beyond simple heatmap visualization, this thesis offers a step toward the explainable AI that pathologists require before such models can be confidently integrated into clinical practice.

Use of AI in thesis

Artificial intelligence (AI) tools were used in a limited and supportive capacity during the preparation of this thesis.

- Elicit was used at the initial stage of the research process to assist in searching for relevant academic literature and to obtain an overview of the research topic. All identified sources were critically evaluated and selected by the author.
- ChatGPT was used during the project to provide support in code understanding, development, and debugging. The primary codebase used in this thesis was provided by the Bioimage Informatics Group (from previous similar tasks) or machine learning coursework. All code was adapted, verified, and validated by the author to ensure its correctness and suitability for the specific research objectives.
- Grammarly was employed to improve the clarity and quality of the written text, including spelling, and grammatical accuracy.

All intellectual contributions, including the research design, data analysis, interpretation of results, and conclusions, were carried out independently by the author. The use of AI tools was limited to supportive purposes and did not replace critical thinking or academic judgment.

Acknowledgements

I would like to express my sincere gratitude to my supervisors. First and foremost, I am deeply grateful to Professor Pekka Ruusuvuori, Principal Investigator and group leader of the Bioimage Informatics Group at the Institute of Biomedicine, University of Turku, for providing me with this research opportunity. His supervision in the technical aspects, along with the provision of the research topic and a supportive working environment, has been essential to the completion of this thesis. I would also like to extend my heartfelt thanks to Docent Camilla Böckelman, from the University of Helsinki and Helsinki University Hospital, for her clinical expertise in colorectal cancer and for her detailed and insightful feedback, which significantly strengthened the quality of my work. I would also like to extend my sincere thanks to the Department of Pathology at Helsinki University Hospital, and Professor Caj Haglund and Professor Jaana Hagström for providing access to the data utilised in this thesis.

My special thanks go to Amanda, whose work laid the foundation for this thesis. She developed the model, processed the images, and generated the feature vectors that served as the core material for my study. Beyond that, she has been a wonderful collaborator and a great friend throughout this journey. I am also grateful to all members of the Bioimage Informatics Group at the Institute of Biomedicine for their support and willingness to help whenever I had questions, particularly in the field of deep learning.

My Master's degree was a journey filled not only with intensive studying but also with many memorable moments and, most importantly, meaningful friendships. I would like to thank Professor Riku Klén, Professor Diana Toivola, Kari Vienola (Academy Research Fellow), and the program coordinators Nina Vainikka and Anna Jalo for their continuous support from the very first day I started the program.

Finally, my deepest gratitude goes to my family. I am profoundly thankful to my husband for always accompanying me on this journey from Vietnam to Finland, for taking care of our children and managing the household, and for creating the conditions that allowed me to complete my studies and this thesis. I am also deeply grateful to my parents for their endless encouragement and emotional support, even from afar, through countless video calls despite the distance between us.

References

[1] "Colorectal cancer," World Health Organization. Accessed: Jun. 02, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/colorectal-cancer

[2] "Global Cancer Observatory: Cancer today – Data visualization (pie chart)," World Health Organization. Accessed: Mar. 01, 2026. [Online]. Available: https://gco.iarc.fr/today/en/dataviz/pie

[3] J. Verma *et al.*, "From slides to insights: Harnessing deep learning for prognostic survival prediction in human colorectal cancer histology.," *Open Life Sci*, vol. 18, 2023, doi: 10.1515/biol-2022-0777.

[4] X. Jiang *et al.*, "End-to-end prognostication in colorectal cancer by deep learning: a retrospective, multicentre study," *The Lancet Digital Health*, vol. 6, no. 1, pp. e33–e43, Jan. 2024, doi: 10.1016/S2589-7500(23)00208-X.

[5] O. Marzouk and J. Schofield, "Review of Histopathological and Molecular Prognostic Features in Colorectal Cancer," *Cancers*, vol. 3, no. 2, pp. 2767–2810, 2011, doi: 10.3390/cancers3022767.

[6] M. Capuozzo *et al.*, "Genetic, Epidemiological, Clinical, and Therapeutic Trajectories in Colon and Rectal Cancers," *Cancers*, vol. 17, 2025, [Online]. Available: https://api.semanticscholar.org/CorpusID:282667385

[7] G. Puppa, A. Sonzogni, R. Colombari, and G. Pelosi, "TNM Staging System of Colorectal Carcinoma: A Critical Appraisal of Challenging Issues," *Archives of Pathology & Laboratory Medicine*, vol. 134, no. 6, pp. 837–852, Jun. 2010, doi: 10.5858/134.6.837.

[8] M. E. A. Elforaici *et al.*, "Semi-supervised ViT knowledge distillation network with style transfer normalization for colorectal liver metastases survival prediction.," *Med Image Anal*, vol. 99, p. 103346, Jan. 2025, doi: 10.1016/j.media.2024.103346.

[9] B. Wei, L. Li, Y. Feng, S. Liu, P. Fu, and L. Tian, "Exploring prognostic biomarkers in pathological images of colorectal cancer patients via deep learning.," *J Pathol Clin Res*, vol. 10, no. 6, p. e70003, Nov. 2024, doi: 10.1002/2056-4538.70003.

[10] E. Wulczyn *et al.*, "Interpretable survival prediction for colorectal cancer using deep learning," *npj Digital Medicine*, vol. 4, no. 1, p. 71, Apr. 2021, doi: 10.1038/s41746-021-00427-2.

[11] D. Bychkov *et al.*, “Deep learning based tissue analysis predicts outcome in colorectal cancer,” *Scientific Reports*, vol. 8, no. 1, p. 3395, Feb. 2018, doi: 10.1038/s41598-018-21758-3.

[12] O.-J. Skrede *et al.*, “Deep learning for prediction of colorectal cancer outcome: a discovery and validation study,” *The Lancet*, vol. 395, no. 10221, pp. 350–360, Feb. 2020, doi: 10.1016/S0140-6736(19)32998-8.

[13] M. Y. Lu, D. F. K. 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, Jun. 2021, doi: 10.1038/s41551-020-00682-w.

[14] X. Wang *et al.*, “Foundation Model for Predicting Prognosis and Adjuvant Therapy Benefit From Digital Pathology in GI Cancers.,” *J Clin Oncol*, p. JCO2401501, Apr. 2025, doi: 10.1200/JCO-24-01501.

[15] R. J. Chen *et al.*, “Towards a general-purpose foundation model for computational pathology,” *Nature Medicine*, vol. 30, no. 3, pp. 850–862, Mar. 2024, doi: 10.1038/s41591-024-02857-3.

[16] U. Sajjad *et al.*, “Morphology-Aware Prognostic model for Five-Year Survival Prediction in Colorectal Cancer from H&E Whole Slide Images,” *ArXiv*, vol. abs/2510.14800, 2025, [Online]. Available: https://api.semanticscholar.org/CorpusID:282138563

[17] S. Wu, Y. Zhang, Z. Lin, and M. Wei, “Global burden of colorectal cancer in 2022 and projections to 2050: incidence and mortality estimates from GLOBOCAN,” *BMC Cancer*, vol. 25, no. 1, p. 1770, Nov. 2025, doi: 10.1186/s12885-025-15138-0.

[18] N. Keum and E. Giovannucci, “Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies,” *Nature Reviews Gastroenterology & Hepatology*, vol. 16, no. 12, pp. 713–732, Dec. 2019, doi: 10.1038/s41575-019-0189-8.

[19] K. Chen, G. Collins, H. Wang, and J. W. T. Toh, “Pathological Features and Prognostication in Colorectal Cancer.,” *Curr Oncol*, vol. 28, no. 6, pp. 5356–5383, Dec. 2021, doi: 10.3390/curroncol28060447.

[20] W. Jiang *et al.*, “Multimodal tumor microenvironment signature of colorectal cancer for prediction prognosis and chemotherapy benefit.,” *NPJ Precis Oncol*, vol. 9, no. 1, p. 270, Aug. 2025, doi: 10.1038/s41698-025-01069-3.

[21] M. J. Kim *et al.*, “Survival Paradox Between Stage IIB/C (T4N0) and Stage IIIA (T1-2N1) Colon Cancer,” *Annals of Surgical Oncology*, vol. 22, no. 2, pp. 505–512, Feb. 2015, doi: 10.1245/s10434-014-3982-1.

[22] H. T. Cotan *et al.*, “Prognostic and Predictive Determinants of Colorectal Cancer: A Comprehensive Review.,” *Cancers (Basel)*, vol. 16, no. 23, Nov. 2024, doi: 10.3390/cancers16233928.

[23] R. P. Graham *et al.*, “Tumor Budding in Colorectal Carcinoma: Confirmation of Prognostic Significance and Histologic Cutoff in a Population-based Cohort.,” *Am J Surg Pathol*, vol. 39, no. 10, pp. 1340–1346, Oct. 2015, doi: 10.1097/PAS.0000000000000504.

[24] “TNM staging for bowel cancer.” Accessed: Jun. 25, 2025. [Online]. Available: https://www.cancerresearchuk.org/about-cancer/bowel-cancer/stages-types-and-grades/TNM-staging

[25] D. M. Karamchandani *et al.*, “Challenges with colorectal cancer staging: results of an international study,” *Modern Pathology*, vol. 33, no. 1, pp. 153–163, Jan. 2020, doi: 10.1038/s41379-019-0344-3.

[26] M. Urbanowicz, H. I. Grabsch, F. Fiteni, Y. Liu, C. Caballero, and J.-F. Fléjou, “An international survey-based study on colorectal cancer pathology reporting-guidelines versus local practice.,” *Virchows Arch*, vol. 473, no. 6, pp. 697–708, Dec. 2018, doi: 10.1007/s00428-018-2457-3.

[27] C. E. L. Klaver *et al.*, “Interobserver, intraobserver, and interlaboratory variability in reporting pT4a colon cancer.,” *Virchows Arch*, vol. 476, no. 2, pp. 219–230, Feb. 2020, doi: 10.1007/s00428-019-02663-0.

[28] A. Davri *et al.*, “Deep Learning on Histopathological Images for Colorectal Cancer Diagnosis: A Systematic Review.,” *Diagnostics (Basel)*, vol. 12, no. 4, Mar. 2022, doi: 10.3390/diagnostics12040837.

[29] J. van der Laak, G. Litjens, and F. Ciompi, “Deep learning in histopathology: the path to the clinic,” *Nature Medicine*, vol. 27, no. 5, pp. 775–784, May 2021, doi: 10.1038/s41591-021-01343-4.

[30] B. Acs, M. Rantalainen, and J. Hartman, “Artificial intelligence as the next step towards precision pathology.,” *J Intern Med*, vol. 288, no. 1, pp. 62–81, Jul. 2020, doi: 10.1111/joim.13030.

[31] S. Huang, J. Yang, S. Fong, and Q. Zhao, “Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges,” *Cancer Letters*, vol. 471, pp. 61–71, 2020, doi: https://doi.org/10.1016/j.canlet.2019.12.007.

[32] C. McGenity *et al.*, “Artificial intelligence in digital pathology: a systematic review and meta-analysis of diagnostic test accuracy.,” *NPJ Digit Med*, vol. 7, no. 1, p. 114, May 2024, doi: 10.1038/s41746-024-01106-8.

[33] S. Schiele *et al.*, “Deep Learning Prediction of Metastasis in Locally Advanced Colon Cancer Using Binary Histologic Tumor Images,” *Cancers*, vol. 13, no. 9, 2021, doi: 10.3390/cancers13092074.

[34] S. Foersch *et al.*, “Multistain deep learning for prediction of prognosis and therapy response in colorectal cancer,” *Nature Medicine*, vol. 29, no. 2, pp. 430–439, Feb. 2023, doi: 10.1038/s41591-022-02134-1.

[35] J. N. Kather *et al.*, “Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study.,” *PLoS Med*, vol. 16, no. 1, p. e1002730, Jan. 2019, doi: 10.1371/journal.pmed.1002730.

[36] O. G. F. Geessink *et al.*, “Computer aided quantification of intratumoral stroma yields an independent prognosticator in rectal cancer,” *Cellular Oncology*, vol. 42, no. 3, pp. 331–341, Jun. 2019, doi: 10.1007/s13402-019-00429-z.

[37] C. Sun *et al.*, “Deep learning with whole slide images can improve the prognostic risk stratification with stage III colorectal cancer,” *Computer Methods and Programs in Biomedicine*, vol. 221, p. 106914, 2022, doi: https://doi.org/10.1016/j.cmpb.2022.106914.

[38] L. Qu, Y. Ma, X. Luo, Q. Guo, M. Wang, and Z. Song, “Rethinking multiple instance learning for whole slide image classification: A good instance classifier is all you need,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[39] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, “Data-efficient and weakly supervised computational pathology on whole-slide images.,” *Nat Biomed Eng*, vol. 5, no. 6, pp. 555–570, Jun. 2021, doi: 10.1038/s41551-020-00682-w.

[40] R. J. Chen *et al.*, “Pan-cancer integrative histology-genomic analysis via multimodal deep learning,” *Cancer Cell*, vol. 40, no. 8, pp. 865-878.e6, Aug. 2022, doi: 10.1016/j.ccell.2022.07.004.

[41] J. Zhou *et al.*, “Integrative deep learning analysis improves colon adenocarcinoma patient stratification at risk for mortality.,” *EBioMedicine*, vol. 94, p. 104726, Aug. 2023, doi: 10.1016/j.ebiom.2023.104726.

[42] O. Ogundipe, Z. Kurt, and W. L. Woo, “Deep neural networks integrating genomics and histopathological images for predicting stages and survival time-to-event in colon cancer.,” *PLoS One*, vol. 19, no. 9, p. e0305268, 2024, doi: 10.1371/journal.pone.0305268.

[43] T. Ding *et al.*, “A multimodal whole-slide foundation model for pathology,” *Nature Medicine*, vol. 31, no. 11, pp. 3749–3761, Nov. 2025, doi: 10.1038/s41591-025-03982-3.

[44] S. Lou *et al.*, “Development and validation of a deep learning-based pathomics signature for prognosis and chemotherapy benefits in colorectal cancer: a retrospective multicenter cohort study.,” *Front Immunol*, vol. 16, p. 1602909, 2025, doi: 10.3389/fimmu.2025.1602909.

[45] F. Prezja *et al.*, “Improved accuracy in colorectal cancer tissue decomposition through refinement of established deep learning solutions,” *Scientific Reports*, vol. 13, no. 1, p. 15879, Sep. 2023, doi: 10.1038/s41598-023-42357-x.

[46] X. Li, J. Jonnagaddala, M.-Y. Cen, H. Zhang, and X. Xu, “Colorectal Cancer Survival Prediction Using Deep Distribution Based Multiple-Instance Learning,” *Entropy*, vol. 24, 2022, [Online]. Available: https://api.semanticscholar.org/CorpusID:248377142

[47] ∗ Christian Abbet, ∗ Linda Studer, I. Zlobec, and J.-P. Thiran, “Toward Automatic Tumor-Stroma Ratio Assessment for Survival Analysis in Colorectal Cancer,” [Online]. Available: https://api.semanticscholar.org/CorpusID:268291377

[48] T. P. T. Armand, S. Bhattacharjee, K. A. Nfor, and H.-C. Kim, “Automated tumor stroma ratio assessment in colorectal cancer using hybrid deep learning approach,” *Scientific Reports*, vol. 15, no. 1, p. 40927, Nov. 2025, doi: 10.1038/s41598-025-24229-8.

[49] Z. Xu *et al.*, “A deep learning quantified stroma-immune score to predict survival of patients with stage II–III colorectal cancer,” *Cancer Cell International*, vol. 21, no. 1, p. 585, Oct. 2021, doi: 10.1186/s12935-021-02297-w.

[50] H. Xu *et al.*, “Spatial analysis of tumor-infiltrating lymphocytes in histological sections using deep learning techniques predicts survival in colorectal carcinoma.,” *J Pathol Clin Res*, vol. 8, no. 4, pp. 327–339, Jul. 2022, doi: 10.1002/cjp2.273.

[51] X. Xiao, Z. Wang, Y. Kong, and H. Lu, “Deep learning-based morphological feature analysis and the prognostic association study in colon adenocarcinoma histopathological images.,” *Front Oncol*, vol. 13, p. 1081529, 2023, doi: 10.3389/fonc.2023.1081529.

[52] A. Laurila, “Foundation AI model for colorectal cancer prognosis : Development of a Pipeline Predicting 5-Year Outcomes from Histopathology Images,” Tampere University, 2025. [Online]. Available: https://trepo.tuni.fi/handle/10024/232897

[53] M. Khened, A. Kori, H. Rajkumar, G. Krishnamurthi, and B. Srinivasan, “A generalized deep learning framework for whole-slide image segmentation and analysis,” *Scientific Reports*, vol. 11, no. 1, p. 11579, Jun. 2021, doi: 10.1038/s41598-021-90444-8.

[54] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 3rd ed. Upper Saddle River, New Jersey, USA: Pearson Prentice Hall, 2008.

[55] U. Khan *et al.*, “Staining normalization in histopathology: Method benchmarking using multicenter dataset,” *Scientific Reports*, Feb. 2026, doi: 10.1038/s41598-026-40943-3.

[56] M. Macenko *et al.*, “A method for normalizing histology slides for quantitative analysis,” in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, Jul. 2009, pp. 1107–1110. doi: 10.1109/ISBI.2009.5193250.

[57] Mahmood Lab, “UNI2-h: Universal Histopathology Foundation Model,” Hugging Face. [Online]. Available: https://huggingface.co/MahmoodLab/UNI2-h

[58] H. Wang *et al.*, “Rethinking Multiple Instance Learning for Whole Slide Image Classification: A Bag-Level Classifier is a Good Instance-Level Teacher.,” *IEEE Trans Med Imaging*, vol. 43, no. 11, pp. 3964–3976, Nov. 2024, doi: 10.1109/TMI.2024.3404549.

[59] L. Cai, S. Huang, Y. Zhang, J. Lu, and Y. Zhang, “Rethinking attention-based multiple instance learning for whole-slide pathological image classification: An instance attribute viewpoint,” *arXiv preprint arXiv:2404.00351*, 2024.

[60] A. K. Jain, “Data clustering: 50 years beyond K-means,” *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, Jun. 2010, doi: 10.1016/j.patrec.2009.09.011.

[61] A. Jain, “Unsupervised Learning: K-Means Clustering,” Towards Data Science. Accessed: Dec. 20, 2025. [Online]. Available: https://towardsdatascience.com/unsupervised-learning-k-means-clustering-27416b95af27/

[62] I. P. Carrascosa, “K-Means Cluster Evaluation with Silhouette Analysis,” Machine Learning Mastery. Accessed: Dec. 20, 2025. [Online]. Available: https://machinelearningmastery.com/k-means-cluster-evaluation-with-silhouette-analysis/

[63] M. Z. I. Chowdhury and T. C. Turin, “Variable selection strategies and its importance in clinical prediction modelling.,” *Fam Med Community Health*, vol. 8, no. 1, p. e000262, 2020, doi: 10.1136/fmch-2019-000262.

[64] C. J. A. Punt *et al.*, “Endpoints in adjuvant treatment trials: a systematic review of the literature in colon cancer and proposed definitions for future trials.,” *J Natl Cancer Inst*, vol. 99, no. 13, pp. 998–1003, Jul. 2007, doi: 10.1093/jnci/djm024.

[65] “F1_score,” Scikit-learn Machine Learning in Python. Accessed: Nov. 12, 2025. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

[66] “What is Harrell’s C-index?,” Statistical Odds & Ends. Accessed: Oct. 12, 2025. [Online]. Available: https://statisticaloddsandends.wordpress.com/2019/10/26/what-is-harrells-c-index/

[67] L. van der Maaten and G. Hinton, “Visualizing Data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

[68] S. Lin, H. Zhou, M. Watson, R. Govindan, R. J. Cote, and C. Yang, “Impact of stain variation and color normalization for prognostic predictions in pathology,” *Scientific Reports*, vol. 15, no. 1, p. 2369, Jan. 2025, doi: 10.1038/s41598-024-83267-w.

[69] Md. Z. Hoque, A. Keskinarkaus, P. Nyberg, and T. Seppänen, “Stain normalization methods for histopathology image analysis: A comprehensive review and experimental comparison,” *Information Fusion*, vol. 102, p. 101997, Feb. 2024, doi: 10.1016/j.inffus.2023.101997.

[70] N. Hegde *et al.*, “Similar image search for histopathology: SMILY,” *npj Digital Medicine*, vol. 2, no. 1, p. 56, Jun. 2019, doi: 10.1038/s41746-019-0131-z.

[71] H. Ueno *et al.*, “Prognostic value of desmoplastic reaction characterisation in stage II colon cancer: prospective validation in a Phase 3 study (SACURA Trial),” *British Journal of Cancer*, vol. 124, no. 6, pp. 1088–1097, Mar. 2021, doi: 10.1038/s41416-020-01222-8.

[72] N. Akimoto *et al.*, “Desmoplastic Reaction, Immune Cell Response, and Prognosis in Colorectal Cancer,” *Frontiers in Immunology*, vol. Volume 13-2022, 2022, doi: 10.3389/fimmu.2022.840198.

[73] Q. Hu *et al.*, “Desmoplastic Reaction Associates with Prognosis and Adjuvant Chemotherapy Response in Colorectal Cancer: A Multicenter Retrospective Study.,” *Cancer Res Commun*, vol. 3, no. 6, pp. 1057–1066, Jun. 2023, doi: 10.1158/2767-9764.CRC-23-0073.

[74] A. K. Roseweir, D. C. McMillan, P. G. Horgan, and J. Edwards, “Colorectal cancer subtypes: Translation to routine clinical pathology.,” *Cancer Treat Rev*, vol. 57, pp. 1–7, Jun. 2017, doi: 10.1016/j.ctrv.2017.04.006.

[75] Z. Song *et al.*, “Automatic deep learning-based colorectal adenoma detection system and its similarities with pathologists.,” *BMJ Open*, vol. 10, no. 9, p. e036423, Sep. 2020, doi: 10.1136/bmjopen-2019-036423.