Deep Learning-Driven Diagnostic and Prognostic Solutions for Histopathological Images of Bladder Cancer

by

Saul Fuster Navarro

Thesis submitted in fulfillment of
the requirements for the degree of

PHILOSOPHIAE DOCTOR
(PhD)

University of
Stavanger

Faculty of Science and Technology
Department of Electrical Engineering and Computer Science
2024

University of Stavanger
N-4036 Stavanger
NORWAY
www.uis.no

ISBN 978-82-8439-241-7
ISSN 1890-1387

PhD Thesis UiS no. 763

Preface

This thesis is submitted as partial fulfilment of the requirements for the degree of *Philosophiae Doctor* at the University of Stavanger, Norway. The research has been carried out at the Department of Electrical Engineering and Computer Science, University of Stavanger in the period of September 2020 to February 2024. The compulsory courses were taken at the University of Stavanger.

The thesis is based on a collection of five papers - three published and two currently under review. For increased readability, the papers have been reformatted for alignment with the format of the thesis and are included as chapters.

Saul Fuster Navarro, May 2024

Abstract

This thesis presents a comprehensive investigation into the development and application of advanced computational techniques for the extraction of crucial diagnostic and prognostic information from histological images of non-muscle invasive bladder cancer (NMIBC). Computational pathology (CPATH) relies on digitized high-resolution tissue samples, referred to as whole slide images (WSIs). Histological examination of WSIs plays a pivotal role in the diagnosis and prognosis of NMIBC. The primary focus of this research is the utilization of deep learning algorithms to automatically analyze histological images and extract visual cues with diagnostic and prognostic significance.

With respect to diagnostics, several convolutional neural network architectures are designed and trained on diverse datasets of NMIBC tissue specimens to identify and classify key histological features, including tumor grading and staging. Moreover, the variability of histological visual features between pathology laboratories during the training of convolutional neural network (CNN) models is questioned. Emphasis is placed on the development of label-efficient guidelines for domain-adapting deep learning models. In addition, an architecture for machine learning is introduced to stratify regions of interest (ROIs) in weakly supervised learning. This additional data stratification aids in localizing ROIs and mitigating cross-noise variability among them.

Furthermore, this thesis explores the integration of deep learning techniques for prognostic assessment. Through an analysis of the relative spatial distribution among urothelium and contingent stromal immune cells, our model predicts patient treatment outcomes and the likelihood of recurrence with a high degree of precision. These prognostic models provide invaluable support to clinicians in customizing personalized treatment strategies and offering patient counseling.

The work presented in this thesis represents a substantial advancement toward improving the diagnostic and prognostic capabilities in the management of NMIBC. Leveraging the potential of computational analysis, we offer pathologists state-of-the-art tools to augment diagnostic precision and optimize patient care, ultimately contributing to better outcomes and quality of life for individuals affected by this prevalent form of bladder cancer.

Acknowledgements

First and foremost, I extend my deepest appreciation to my supervisor, Professor Kjersti Engan, whose guidance, expertise, and support have been exceptional throughout this journey. We've encountered numerous challenges and triumphs on this journey, and she has consistently stood by my side. For all the eye-opening brainstorming sessions in her office. I feel fortunate and could not have wished for a more extraordinary supervisor. I am thankful to my co-supervisor, Professor Trygve Eftestøl, for his insightful feedback and constructive criticism, which have greatly enhanced the quality of this work. I remember vividly the first day of my PhD, I was told: "*You have been lucky to have such amazing supervisors*". Looking back, they were completely right. I want to thank specially the Head of the Department Dr. Tom Ryen for creating an incredible academic environment. I want to acknowledge the support and resources provided by the University of Stavanger and the Department of Electrical Engineering and Computer Science. All have been essential to the completion of this dissertation.

My most sincere gratitude to all colleagues at IDE with which I share many memories. Having lunch, talking in the corridors, invading office KE E-401 out of nowhere, going for a coffee in the Bokkaféen, gym sessions, a BBQ near the botanic garden, a beer at Cardinal, playing board games, birthday cake in the lunchroom, and countless others. You made me feel deeply appreciated and valued.

I would like to commend the CLARIFY network for the exceptional individuals I've had the privilege of meeting and the meaningful experiences we've shared. Special thanks to all ESRs, I learned from them personally and professionally. I thoroughly enjoyed our training schools and getting to know them closely. In Granada walking around the old town, in Amsterdam going for a boat trip, in Stavanger hiking Preikestolen and a lovely farewell in Valencia.

I can never forget all friends back in La Pobla de Vallbona. Despite being away in Norway for over three years, whenever I'm back, it feels like I never left. I always look forward to meet them at the bar for esmorzaret, cremaet and chatting for hours.

I would like to extend a peculiar thanks to my feline companions, Sherlock and Watson, who, with their comforting purrs and playful antics, were a source of joy that helped me maintain my sanity throughout this journey.

I cannot praise enough my parents and my brother for being encouraging, supportive, understanding, present regardless of the distance and loving most of all. To all my family in La Pobla de Vallbona, Puertollano and Barcelona, who always have cared and hoped the best for me. To my in-laws, whose emotional and professional guidance has been invaluable.

All that I've expressed so far pales in comparison to the depth of my love for the most important person in my life, Jen. She accompanies me every day to help me grow, both personally and professionally. Thank you for making me a better version of myself, for sharing your madness and joy with me. Thank you for being a source of inspiration and for always being by my side, supporting me in the tough times and celebrating with me in the good ones. Thank you for the life you give me.

Finally, I dedicate this work to la yaya, whose memory continues to inspire me.

Thank you, one and all, for being a part of this significant chapter in my life. I cannot thank you enough for all your support.

Saul Fuster Navarro, May 2024

List of Notations

(Some symbols have different meanings in different chapters. Hence, they appear in multiple chapters of this list.)

Chapter 2 - Technical Background

$\mathcal{D}$	Dataset
x	Data point
y	Label
M	Machine learning model
$\hat{y}$	Output prediction
$\mathcal{L}$	Loss function
i, j, k	Index
x_i	Neuron i input
w_i	Neuron i weight
b	Neuron bias
θ	Activation function
N	Number of samples in a dataset $\mathcal{D}$
$\mathcal{X}$	Bag of instances
l	Instance number
L	Number of instances
$\mathcal{X}_\mathcal{T}$	Data points of training set

N_T	Number of samples in training set
$\mathcal{X}_l$	Data points of initial training set
N_0	Number of samples in initial training set
$\mathcal{P}$	Data points of data pools
N_P	Number of samples in data pools
N_b	Number of samples in a batch
$\mathcal{L}_{\mathbf{c}}$	Contrastive loss
$\mathcal{L}_{\mathbf{ce}}$	Cross entropy loss
α, β, γ	FTL parameter
$\hat{y}_{i,c}$	Probability of a data point i to be as class c in $\hat{y}$
$\hat{y}_{i,\hat{c}}$	Probability of a data point i to not be as class c
$\text{sim}(\mathbf{z}_i, \mathbf{z}_j)$	Cosine similarity between feature vectors $\mathbf{z}_i$ and $\mathbf{z}_j$
$1_{[k \neq i]}$	Binary indicator to indicate all other instances than i
τ	Temperature coefficient
$P_{(i)}$	Cardinality of i
$\mathcal{I}$	Image
$x_{\mathbf{h}}$	Horizontal coordinates of $\mathcal{I}$
$y_{\mathbf{v}}$	Vertical coordinates of $\mathcal{I}$
$\mathcal{K}$	Kernel
$\circledast$	Convolution
$\widetilde{k_1}$	Half-height of the kernel $\mathcal{K}$
$\widetilde{k_2}$	Half-width of the kernel $\mathcal{K}$
a_i	Attention score for instance with index i
$\odot$	Element-wise multiplication
$\tanh(\cdot)$	Hyperbolic tangent activation
$\text{sigm}(\cdot)$	Sigmoid activation

M	Number of neurons in intermediate layer of attention mechanism
$\mathbf{w}$	Trainable parameters with shape $\mathbb{R}^{L\times 1}$
$\mathbf{V}$	Trainable parameters with shape $\mathbb{R}^{L\times M}$
$\mathbf{U}$	Trainable parameters with shape $\mathbb{R}^{L\times M}$

Chapter 5 - Learning Methods

$\mathcal{X}$	Training set
$\mathcal{X}_0$	Initial training subset from $\mathcal{X}$
$\mathcal{P}$	Pools of data subset from $\mathcal{X}$
$\mathcal{V}$	Validation set
i	Class index
J	AL iteration index
S	Sample size per iteration
N_j^i	Number of samples from class i in iteration j
δ_i^j	Relative class i distribution in iteration j
$\mathcal{X}_j$	Training subset at iteration j
$\mathcal{X}_j^i$	Class i subset for $\mathcal{X}_j$
G_f	CNN feature extractor
$\mathcal{X}, \mathcal{Y}$	Dataset pair of bags of instances and labels
i	Bag index
N	Number of pairs in dataset $\mathcal{X}, \mathcal{Y}$
$\mathbf{X}^i$	Bag of instances i from dataset $\mathcal{X}$
$\mathbf{y}^i$	Label i from dataset $\mathcal{Y}$
l	Instance index
L	Number of instances in bag $\mathbf{X}$
$\mathbf{x}_l$	Instance l from bag $\mathbf{X}$

G_f	CNN feature extractor
$\bar{\mathbf{x}}$	Feature embedding representation of $\mathbf{x}$
Ξ	Aggregation function for mean and max operators
$\hat{y}$	Model output prediction
Θ_c	Classifier
j	Index of nesting level
J	Number of nesting levels
$\mathbf{X}_j$	Set of instances or bag at level j
k	Index of inner-bag
K	Number of inner-bags
$\mathbf{X}_{j,k}$	Inner-bag k at level j
K_j	Number of bags at level j
$L_{j,k}$	Number of instances or bags in a bag j, k
$\mathbf{x}_{j,k,l}$	Instance l in inner-bag $\mathbf{X}_{j,k}$
y_l^j	Latent label of an instance l in a level j, omitting bag index k
δ	Index compression for (j, k, l)
δ'	Index compression for $(j+1, k', l')$
γ	Index compression for (j, k)
a_δ	Attention score for instance with index δ
$\tilde{a}_\delta$	Normalized attention score for instance with index δ

Paper I - Nested Multiple Instance Learning with Attention Mechanisms

$\mathcal{X}, \mathcal{Y}$	Dataset pair of bags of instances and labels
i	Bag index
N	Number of pairs in dataset $\mathcal{X}, \mathcal{Y}$
$\mathbf{X}^i$	Bag of instances i from dataset $\mathcal{X}$
$\mathbf{y}^i$	Label i from dataset $\mathcal{Y}$
l	Instance index
L	Number of instances in bag $\mathbf{X}$
$\mathbf{x}_l$	Instance l from bag $\mathbf{X}$
G_f	CNN feature extractor
$\bar{\mathbf{x}}$	Feature embedding representation of $\mathbf{x}$
Ξ	Aggregation function for mean and max operators
$\hat{y}$	Model output prediction
Θ_c	Classifier
j	Index of nesting level
J	Number of nesting levels
$\mathbf{X}_j$	Set of instances or bag at level j
k	Index of inner-bag
K	Number of inner-bags
$\mathbf{X}_{j,k}$	Inner-bag k at level j
K_j	Number of bags at level j
$L_{j,k}$	Number of instances or bags in a bag j, k
$\mathbf{x}_{j,k,l}$	Instance l in inner-bag $\mathbf{X}_{j,k}$
y_l^j	Latent label of an instance l in a level j, omitting bag index k
δ	Index compression for (j, k, l)

δ'	Index compression for $(j+1, k', l')$
γ	Index compression for (j, k)
a_δ	Attention score for instance with index δ
$\tilde{a}_\delta$	Normalized attention score for instance with index δ
$\odot$	Element-wise multiplication
$\tanh(\cdot)$	Hyperbolic tangent activation
$\text{sigm}(\cdot)$	Sigmoid activation
M	Number of neurons in intermediate layer of attention mechanism
$\mathbf{w}$	Trainable parameters with shape $\mathbb{R}^{L\times 1}$
$\mathbf{V}$	Trainable parameters with shape $\mathbb{R}^{L\times M}$
$\mathbf{U}$	Trainable parameters with shape $\mathbb{R}^{L\times M}$

Paper II - Active Learning Based Domain Adaptation for Tissue Segmentation of Histopathological Images

$\mathcal{X}$	Training set
$\mathcal{X}_0$	Initial training subset from $\mathcal{X}$
$\mathcal{P}$	Pools of data subset from $\mathcal{X}$
$\mathcal{V}$	Validation set
i	Class index
J	AL iteration index
S	Sample size per iteration
N_j^i	Number of samples from class i in iteration j
δ_i^j	Relative class i distribution in iteration j
$\mathcal{X}_j$	Training subset at iteration j
$\mathcal{X}_j^i$	Class i subset for $\mathcal{X}_j$

TRI – AL$_{\mathbf{OOD}}$	Deep learning model trained on a AL manner where a class pool $\mathcal{P}^i$ must exhaust all data points
TRI – AL$_{\mathbf{ENT}}$	Deep learning model trained on a AL query based on entropy uncertainty
TRI – AL$_{\mathbf{ITER}}$	Deep learning model trained on a AL manner with J iterations
TRI – CNN	Deep learning model trained on an external cohort
TRI – SL	Deep learning model trained on a SL manner

Paper III - Invasive Cancerous Area Detection in Non-Muscle Invasive Bladder Cancer Whole Slide Images

E_{mag}	Set of experiments for magnification level exploration
E_{ds}	Set of experiments for artifact tissue type differentiation
E_{emc}	Set of experiments using only patients from EMC cohort
E_{tils}	Set of experiments excluding tiles containing TILs

Paper IV - Advancing Histopathological Bladder Cancer Grading with Weakly Supervised Deep Learning

$\mathcal{T}$	Triplet of three tiles at various magnification levels
$\mathcal{Y}$	Set of classes
N_b	Number of tiles in a blob of connected tiles
$T_{\mathbf{LOWER}}$	Lower threshold
$T_{\mathbf{UPPER}}$	Upper threshold
N_c	Number of clusters
$\mathcal{X}, \mathcal{Y}$	Dataset pair of bags of instances and labels
i	Bag index
N	Number of pairs in dataset $\mathcal{X}, \mathcal{Y}$

$\mathbf{X}^i$	Bag of instances i from dataset $\mathcal{X}$
$\mathbf{y}^i$	Label i from dataset $\mathcal{Y}$
l	Instance index
L	Number of instances in bag $\mathbf{X}$
$\mathbf{x}_l$	Instance l from bag $\mathbf{X}$
G_θ	CNN feature extractor
$\mathbf{h}$	Feature embedding representation of $\mathbf{x}$
$\mathbf{A}$	Attention scores
$\hat{y}$	Model output prediction
Ψ_ρ	Classifier
$\mathbf{H_{WSI}}$	WSI feature embedding representation
$\mathbf{H_{REG}}$	Region feature embedding representation
$\mathbf{h_{TILE}}$	Tile feature embedding representation
μ	Mean value
σ	Standard deviation

Paper V - Self-Contrastive Weakly Supervised Learning Framework for Prognostic Prediction Using Whole Slide Images

$S_{\mathbf{EMC}}$	Erasmus Medical Center cohort
$S_{\mathbf{SUH}}$	Stavanger University Hospital cohort
D_y	Dataset corresponding to a region of interest y
y	Region of interest corresponding to a tissue type
$\mathcal{X}, \mathcal{Y}$	Dataset pair of bags of instances and labels
x	Tile of a WSI, corresponding to $\mathcal{X}$
N	Batch size
$\mathcal{H}$	Feature embeddings of $\mathcal{X}$
$\mathcal{Z}$	Projected features of $\mathcal{H}$

G_θ	Feature extractor
F_ϕ	Multi-layer perceptron
$\text{sim}(\mathbf{z}_i, \mathbf{z}_j)$	Cosine similarity between feature vectors $\mathbf{z}_i$ and $\mathbf{z}_j$
$1_{[k \neq i]}$	Binary indicator to indicate all other instances than i
τ	Temperature coefficient
$P_{(i)}$	Cardinality of i
$\mathcal{L}_c$	Unsupervised contrastive loss
$\mathcal{L}_{sc}$	Supervised contrastive loss
$\mathcal{L}_{multi}$	Multi-task loss
α_c	Scaling factor of $\mathcal{L}_c$
α_{ce}	Scaling factor of $\mathcal{L}_{ce}$
i	Bag index
$\mathbf{H}^i$	Bag of instances i from dataset $\mathcal{H}$
$\mathbf{y}^i$	Label i from dataset $\mathcal{Y}$
l	Instance index
L	Number of instances in bag $\mathbf{H}$
$\mathbf{h}_l$	Instance l from bag $\mathbf{H}$
$\hat{y}$	Model output prediction
Ψ_p	Classifier
a_i	Attention score for instance with index i
$\odot$	Element-wise multiplication
$\tanh(\cdot)$	Hyperbolic tangent activation
$\text{sigm}(\cdot)$	Sigmoid activation
M	Number of neurons in intermediate layer of attention mechanism
$\mathbf{w}$	Trainable parameters with shape $\mathbb{R}^{L \times 1}$

$\mathbf{V}$	Trainable parameters with shape $\mathbb{R}^{L \times M}$
$\mathbf{U}$	Trainable parameters with shape $\mathbb{R}^{L \times M}$
$\mathbf{h_{TILE}}$	Instance-level representation
$\mathbf{H_{REG}}$	Bag of instances belonging to a region
$\mathbf{H_{WSI}}$	Bag of instances belonging to a WSI
k	Region index
K	Number of regions in bag $\mathbf{H}$
$\mathbf{h}_{cli}$	Instance embedding containing clinicopathological information
θ_I	Imagenet weights for G_θ
θ_C	Weights for G_θ trained using $\mathcal{L}_c$
θ_{SC}	Weights for G_θ trained using $\mathcal{L}_{sc}$
θ_{CE}	Weights for G_θ trained using $\mathcal{L}_{ce}$
θ_{MULTI}	Weights for G_θ trained using $\mathcal{L}_{multi}$
α_l, γ_l	Parameters of Focal Tversky Loss
n_b	Number of instances in a batch
lr	Learning rate
opt	Optimizer
d_r	Dropout rate
n_{Θ_ρ}	Number of neurons in Ψ_p
n_{att}	Number of neurons in attention module
$\mathcal{P}_{\sigma,\mu}$	Normal distribution with mean σ and standard deviation μ

List of Abbreviations

AI	Artificial Intelligence
AL	Active Learning
ANNO	Manually annotated region
AUTO	Automatically segmented region
BCG	Bacillus Calmette-Guérin
CAD	Computer-Aided Diagnosis
CIS	Carcinoma In Situ
CNN	Convolutional Neural Network
CPATH	Computational PATHology
CUETO	Club Urológico Español de Tratamiento Oncológico
DI	Di-scale model
DL	Deep Learning
DNN	Deep Neural Network
EAU	European Association of Urology
EMC	Erasmus Medical Center
EORTC	European Organisation for Research and Treatment of Cancer
FNR	False Negative Ratio
FTL	Focal Tversky Loss
GAP	Global Average Pooling

HE	Hematoxylin-Eosin
HES	Hematoxylin-Eosin-Saffron
HR	High-Risk
ICA	Invasive Cancerous Area
LP	Lamina Propria
LSTM	Long Short-Term Memory
MI	Multiple Instance
MIA	Multiple Instance with Attention
MIBC	Muscle Invasive Bladder Cancer
MIL	Multiple Instance Learning
ML	Machine Learning
MLP	Multi Layer Perceptron
NMI	Nested Multiple Instance
NMIA	Nested Multiple Instance with Attention
NMIBC	Non-Muscle Invasive Bladder Cancer
NMIL	Nested Multiple Instance Learning
NMIST	Modified National Institute of Standards and Technology
MONO	Mono-scale model
NN	Neural Network
PCAM	Patch CAMelyon
RNN	Recurrent Neural Network
ROI	Region Of Interest
SGD	Stochastic Gradient Descent
SL	Supervised Learning
SUH	Stavanger University Hospital
T1	Non-invasive papillary carcinoma

Ta	Tumor invades lamina propria
TI	Tversky Index
TIL	Tumour Infiltrating Lymphocyte
TNM	Tumour Node Metastasis
TRI	Tri-scale model
TURBT	TransUrethral Resection of a Bladder Tumour
URO	Urothelium
WHO	World Health Organization
WHO04	2004 World Health Organization classification of papillary urothelial carcinoma
WHO73	1973 World Health Organization classification of papillary urothelial carcinoma
WSI	Whole Slide Image

List of Publications

This dissertation is formed by the main scientific papers that have significantly contributed to the research presented. The following list provides a comprehensive overview of these research papers:

- **Paper 1**

  Nested Multiple Instance Learning with Attention Mechanisms
  S. Fuster, T. Eftestøl, K. Engan
  Published in the Proceedings of the 21st IEEE International Conference on Machine Learning and Applications (ICMLA), 2022

- **Paper 2**

  Active Learning Based Domain Adaptation for Tissue Segmentation of Histopathological Images
  S. Fuster, F. Khoraminia, T. Eftestøl, T. C. M. Zuiverloon, K. Engan
  Published in the Proceedings of the 31th European Signal Processing Conference (EUSIPCO), 2023

- **Paper 3**

  Invasive Cancerous Area Detection in Non-Muscle Invasive Bladder Cancer Whole Slide Images
  S. Fuster, F. Khoraminia, U. Kiraz, N. Kanwal, V. Kvikstad, T. Eftestøl, T. C. M. Zuiverloon, E. A. M. Janssen, K. Engan
  Published in the Proceedings of the IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), 2022

- **Paper 4**

  NMGrad: Advancing Histopathological Bladder Cancer Grading with Weakly Supervised Deep Learning
  S. Fuster, U. Kiraz, T. Eftestøl, E. A. M. Janssen, K. Engan
  Under review

- **Paper 5**

  Self-Contrastive Weakly Supervised Learning Framework for Prognostic Prediction Using Whole Slide Images
  S. Fuster, F. Khoraminia, J. Silva-Rodríguez, U. Kiraz, G. J. L. H. van Leenders, T. Eftestøl, V. Naranjo, E. A. M. Janssen, T. C. M. Zuiverloon, K. Engan
  Under review

Contents

Chapter 1

Introduction

Pathology is a highly specialized field where pathologists analyse biopsies of tissue samples with the main aim to verify cancer diagnosis. Pathological analysis is the core of the diagnostic process. Traditional pathologists' workflow relies on bench marking microscopy assessment based on individual expertise. This expertise is complemented, when needed, with additional literature background or colleagues support, which occasionally leads to a high level of discrepancy among experts in complex scenarios [1]. As a consequence, diagnostic pathology in practice today is still a slow and cumbersome process that relies heavily on the subjective interpretation of a microscopic image. These intrinsic limitations increase when pathologists are not specialized in a particular area, a situation which is frequent in small pathology departments [2]. In these cases, consequently, the probability of early and correct diagnose is hindered, as reproducibility, objectivity and precision are severely limited [3].

In addition to the procedural issues, the workload that pathology departments assume is exponentially growing due to the increasing number of biopsies, cancer cases and screening programs, designed with a very high sensitivity and quite low specificity [4]. This creates an increasing demand to analyse normal or near normal biopsies that it is estimated to go up to 47% by 2040 worldwide [5]. As a result, the saturation of the pathology departments leads to delays in diagnosis that have a significant adverse impact on treatments' assignment and effectiveness [6].

Digital pathology is an image-based information environment that enables the management and interpretation of pathological information generated from the digitalization of a tissue sample [7]. It has been only recently that

improved scanning, storage, data transfer speeds, and advances in software and computer processing power have made digital pathology progress possible. Digital microscopy scanners are today capable of scanning entire stained tissue sections from glass slides in a few seconds and to generate high-resolution digital images from the scanned tissue sections, also named whole slide images (WSI), in a fully automatic way. Thanks to the digitisation of the medical image, non-invasive techniques have appeared to help to the diagnosis through the computational analysis of the image. These new approaches allow the quantification of image biomarkers and the detection and classification of pathology detection by means of artificial intelligence techniques in a fast and efficient way. Computer-aided diagnosis (CAD) systems can support pathologists in the task of diagnosing and giving valuable prognostic information [8].

Bladder cancer (BC) is a heterogeneous disease with high prevalence and recurrence rates [10]. Notably, the stage and grade of bladder cancer plays a very important role in prognostication and risk assessment of this disease, particularly non-muscle invasive bladder cancer (NMIBC). Its clinical behavior is usually related to pathological grade. Low-grade non-invasive papillary carcinoma tumors have a low progression rate and require initial endoscopic treatment and surveillance but rarely present a threat to the patients, while high-grade tumors have a high malignant potential associated with significant progression and cancer death rates.

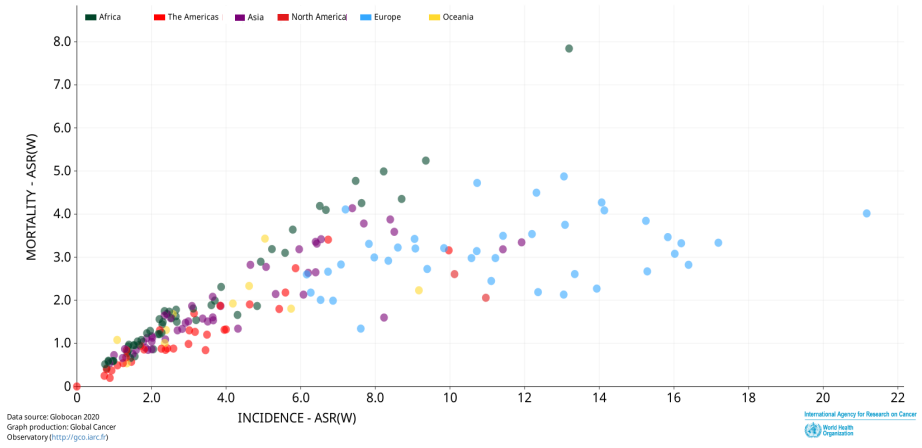


Figure 1.1: Estimated age-standardize mortality rates in 2020 for bladder cancer. Reprinted from Global Cancer Observatory: Cancer Today [9].

Current mortality rates are displayed in Fig. 1.1. On average, 70% of bladder tumors present as superficial disease and the remainder, as muscle-invasive disease. Many characteristics of urothelial carcinoma have been studied in an attempt to predict variable tumor behavior. These include pathologic features, cytologic analysis, and molecular markers. Although the management of NMIBC tumours has significantly improved during the past few years, it remains difficult to predict the heterogeneous outcome of such tumours, especially if high-grade NMIBC is present. Transurethral resection is the initial treatment of choice for NMIBC [11]. However, the high rates of recurrence and significant risk of progression in higher-grade tumours mandate additional therapy with intravesical agents [12]. Accurate staging and grading of the disease is important to decide the optimal treatment. An understanding of the epidemiology helps in the prevention and early detection of the disease.

The 2022 version of the European Association of Urology (EAU) guidelines on non-muscular infiltrating urothelial carcinoma suggests that patients are further stratified into risk groups for low, intermediate, high-risk (HR) and 'highest-risk' based on the hazard to progress to a muscle invasive disease [10]. Accurate assessment of the risk is of utmost importance for NMIBC management, as the treatment strategy depends not only on the presence of muscle invasion. Efficient treatment guidelines would allow uropathologist to distinguish patients who respond to treatment, those who experience recurrence, or progress to a higher stage, from those who do not. This differentiation is particularly relevant at the time of their initial transurethral resection of a bladder tumor (TURBT) [13]. Identifying these patients at the first stage would significantly reduce the mortality rate and treatment cost, hence contributing to more adequate patient-based treatment strategies.

Over the last decade, the field of computational pathology (CPATH) has been extensively explored [14–17]. Among the various cancer types studies in the literature [18–22], research on BC is negligible in comparison [23–25]. Regarding diagnostics, there has been a considerable list of applications, such as tissue segmentation [26–28], cell segmentation [29, 30], grading [31–33] or staging [34–36]. In comparison, prognostic methods have not been researched enough. Some of the most common applications to discern clinical outcome include recurrence of the disease after treatment [32, 37, 38], progression to a higher stage of the disease [38, 39], survival prediction [40, 41] and response to treatment [42, 43]. The implementation of a computer-aided diagnosis (CAD) system in the clinic has the potential to address

several key challenges, including enhancing the reproducibility of diagnoses, reducing variability in interpretations, alleviating the growing workload on healthcare professionals, and ultimately improving both workflow efficiency and patient outcomes.

1.1 Research Challenges and Opportunities

Pathological evaluation of histological features remains the prime method for disease assessment. Despite pathology laboratories slowly shifting towards a fully-digital workflow, there exists a vast amount of manual labour to be performed by pathologists. In the context of large whole slide images (WSIs), this labour mostly consists of carefully visualizing all present tissue areas in the slides and annotating relevant regions of interest (ROIs). Pathology datasets often favor supervised learning for tile-level classification within WSIs [44, 45]. However, at that scale, it is unfeasible to depend on costly and time-consuming manual annotations [46–48]. Moreover, annotation strategies are missing, and efforts towards more efficient guidelines is imperative to reduce excessive input from expert pathologists.

Even if annotated datasets are available, there is a lack of consensus among pathologists with respect to risk assessment for accurate clinical outcome forecasting. These inconsistencies lead to biased or unreliable predictions [49]. While deep learning (DL) supervised learning methods have been widely utilized in the field of pathology, non-supervised approaches for NMIBC are critically uninvestigated. Complementing traditional supervised learning techniques could ultimately enhance diagnostic and prognostic capabilities in NMIBC.

Most of the research literature focuses on diagnostic applications, neglecting the essential verdict of clinical outcome. Although the impact of particular prognostic factors has been explored, they often lack confirmatory evidence that these are intrinsically related to a positive prognosis. To the best of the authors knowledge, no automatic system based on image features for treatment prediction has been reported in the literature.

1.2 Objectives

The main goal of this thesis is to automatically extract diagnostic and prognostic histological features from NMIBC WSIs by means of artificial

intelligence. It is of utmost importance to stratify patients into risk groups to predict clinical outcome, according to EAU guidelines on NMIBC. This risk group categorization is subject to age of the patient, whether it is a primary tumour, tumour size, number of tumours, concomitant carcinoma in situ (CIS), grade and stage. Therefore, algorithms that obtain quantitative results of these factors are necessary. Furthermore, as annotation of WSIs is a cumbersome and time-consuming process, we delve into methodologies aimed at mitigating the reliance on thoroughly annotated data and reducing dependence on expert pathologist opinions.

The thesis objectives are divided into one main objective (O_1) and three sub-objectives (SO_1-SO_3) as follows:

- O_1: Creating clinic-relevant risk factor assessment of NMIBC patients using artificial intelligence.
  - SO_1: Explore the utilization of non-supervised learning approaches to address the absence of annotations.
  - SO_2: Create automated deep learning systems for classifying risk factors based on histological features.
  - SO_3: Propose a deep learning method for predicting clinical outcome from histological slides of NMIBC patients.

1.3 Proposed Approaches

Different ML and DL algorithms and pipelines were proposed in five academic papers to meet the different thesis's objectives. Figure 1.2 displays a general overview of the topics analyzed in the included papers and how they are connected.

The five papers of the thesis focus on the main objective (O_1). The methodology developed in this thesis addresses the first research sub-objective (SO_1): in Paper I, a weakly supervised learning machine learning (ML) architecture is proposed for localization and stratification of the disease. The second research sub-objective SO_2 is addressed in Paper II, III and IV. In Paper II, invasive patterns are discerned from non-invasive using a context-aware multi-magnification convolutional neural network (CNN) model. In Paper III, an active learning approach is proposed for tissue segmentation of BC slides. In Paper IV, an end-to-end DL system is proposed for grading WSIs of NMIBC. The last research sub-objective SO_3

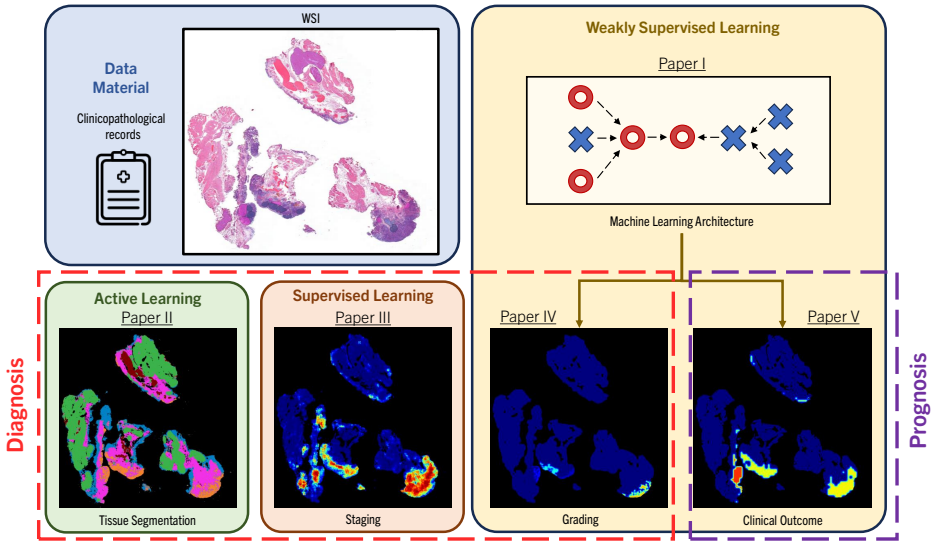


Figure 1.2: Overview of the contributions and their categorization. Non-supervised learning, such as active and weakly supervised, aligns with SO1, diagnosis with SO2, and prognosis with SO3. Paper II, III and IV fall under diagnosis, while Paper V corresponds to prognosis. All Papers but II, explored non-supervised approaches.

is mainly addressed in Paper V, where we concatenate several DL models for extracting tissue relevant areas, learn meaningful features, and predict clinical outcome.

1.4 Project Contributions

The overall contribution of this thesis was to develop algorithms and systems to extract diagnostically relevant information for use with digital pathology, more specifically for histological WSIs of NMIBC. The thesis's main contributions are the five included papers summarized in the following:

I. Saul Fuster, Trygve Eftestøl, Kjersti Engan, **"Nested Multiple Instance Learning with Attention Mechanisms"**, 21st IEEE International Conference on Machine Learning and Applications (ICMLA), 2022.

II. Saul Fuster, Farbod Khoraminia, Trygve Eftestøl, Tahlita C.M. Zuiverloon, Kjersti Engan, **"Active Learning Based Domain Adapta-**

tion for Tissue Segmentation of Histopathological Images", 31st European Signal Processing Conference (EUSIPCO), 2023.

III. Saul Fuster, Farbod Khoraminia, Umay Kiraz, Neel Kanwal, Vebjørn Kvikstad, Trygve Eftestøl, Tahlita C.M. Zuiverloon, Emiel A.M. Janssen, Kjersti Engan, **"Invasive Cancerous Area Detection in Non-Muscle Invasive Bladder Cancer Whole Slide Images"**, IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), 2022.

IV. Saul Fuster, Umay Kiraz, Trygve Eftestøl, Emiel A.M. Janssen, Kjersti Engan, **"NMGrad: Advancing Histopathological Bladder Cancer Grading with Weakly Supervised Deep Learning"**, 2023.

V. Saul Fuster, Farbod Khoraminia, Julio Silva-Rodríguez, Umay Kiraz, Geert J. L. H. van Leenders, Trygve Eftestøl, Valery Naranjo, Emiel A.M. Janssen, Tahlita C.M. Zuiverloon, Kjersti Engan, **"Self-Contrastive Weakly Supervised Learning Framework for Prognostic Prediction Using Whole Slide Images"**, 2023.

Paper I This paper proposes the implementation of a machine learning architecture for weakly supervised learning. The proposed model architecture leverages sets and localization as stratified input. The study was performed to define how intricate sets of data require versatile models.

Paper II This paper proposes a domain adaptation of deep learning models via an active learning procedure. Active learning reduces the number of labeled data necessary in an efficient effort to retrain an algorithm, in comparison to a supervised learning approach. We explore the use of the class balance evolution over the procedure to recommend annotation guidelines.

Paper III This paper explores the detection and localization of invasive cancerous areas in NMIBC WSIs. The models utilized rely on a multiscale input CNN for leveraging cellular features and morphological structures. The research analyzes the strengths, limitations, and potentials of various magnification levels, tissue type differentiation and domain-shift between different hospital cohorts.

Paper IV This paper explores the use of weakly supervised learning techniques for grading NMIBC WSIs. We compare classic approaches to that proposed in Paper I, weighting the impact of magnification levels in addition. Multiple instance learning and its architectural variants showed state of the art performance.

Paper V This paper depicts the use of self-contrastive weakly supervised learning for prognostic applications using histological images. We propose a two-step method for training a feature extractor using self-contrastive learning and multiple instance learning for patient classification, as proposed in Paper I. Furthermore, we delve into the relevance of tissue regions as prognostic markers.

1.5 List of Other Works

In addition to the papers included in this thesis, several other works conducted during the PhD period have contributed to the field, offering alternative perspectives and methodologies that enrich the understanding of deep learning in histopathological image analysis. These publications resulted from collaborations with project partners and supervision of master's students. Contributions involved tasks such as writing and development. The following list presents additional publications related to the thesis topic, distinct from the thesis itself:

VI. Neel Kanwal, Saul Fuster, Farbod Khoraminia, Tahlita C.M. Zuiverloon, Chunming Rong, Kjersti Engan, **"Quantifying the effect of color processing on blood and damaged tissue detection in Whole Slide Images"**, IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), 2022.

VII. Christopher Andreassen, Saul Fuster, Helga Hardardottir, Emiel A.M. Janssen, Kjersti Engan, **"Deep Learning for Predicting Metastasis on Melanoma WSIs"**, IEEE 20th International Symposium on Biomedical Imaging (ISBI), 2023.

VIII. Farbod Khoraminia, Saul Fuster, Neel Kanwal, Mitchell Olislagers, Kjersti Engan, Geert J.L.H. van Leenders, Andrew P. Stubbs, Farhan Akram, Tahlita C.M. Zuiverloon, **"Artificial Intelligence in Digital Pathology for Bladder Cancer: Hype or Hope? A Systematic Review"**, Cancers, 2023.

IX. Marie Bø-Sande, Edvin Benjaminsen, Neel Kanwal, Saul Fuster, Helga Hardardottir, Ingrid Lundal, Emiel A.M. Janssen, Kjersti Engan, **"A Dual Convolutional Neural Network Pipeline for Melanoma Diagnostics and Prognostics"**, Northern Lights Deep Learning (NLDL) Workshop, 2024.

1.6 Outlines

The thesis is structured in four parts: I) *Methodology*, II) *Clinical Applications*, III) *Contributions*, and IV) *Included Papers*.

Methodology describes basics ML and DL concepts relevant for the thesis (Chapter 2). *Clinical Applications* introduces knowledge about WSI and NMIBC, and provides a brief overview of existing approaches to interpret tissue features for diagnostic and prognostic interpretation (Chapter 3). It also provides an overview of the dataset used in the various papers included in the thesis (Chapter 4). *Contributions* summarize the ML and DL methods proposed tfor quantifying risk factors for NMIBC patients (Chapter 5), as well as their application (Chapter 6-8). Finally, it provides remarks on this work (Chapter 9). *Included Papers* lists the included papers in this thesis.

Part I

Methodology

Chapter 2

Technical Background

This chapter introduces the technical background in artificial intelligence and related topics that are the foundation of this thesis.

2.1 Artificial Intelligence

Artificial intelligence (AI) is a field of study aiming to teach intelligent agents to perform human actions. These are computer systems that mimic the decision-making capabilities of a human expert in a specific domain. They use a knowledge base comprising facts and rules to provide expert-level advice or solve complex problems. Some applications may include robotics [50] automatic language replies [51, 52], genetic algorithms [53], among others [54]. The versatility and value of AI is prone to reimagine our society [55–57].

2.1.1 Machine Learning

Machine learning (ML) is a sub-field of AI concerned with developing algorithms without explicit guidance, distilling data patterns. ML was originally categorized into three major subgroups based on the learning paradigm, those groups being supervised, unsupervised and reinforcement learning [58]. Lately, the variance in learning methods has led to the emergence alternative means to exploit the model learning [59, 60].

Datasets, $\mathcal{D}$, are of the essence, as ML relies on data to form algorithms. These are sets of data points coming in various forms, often with a corresponding attribute or label. A dataset contains the insights necessary

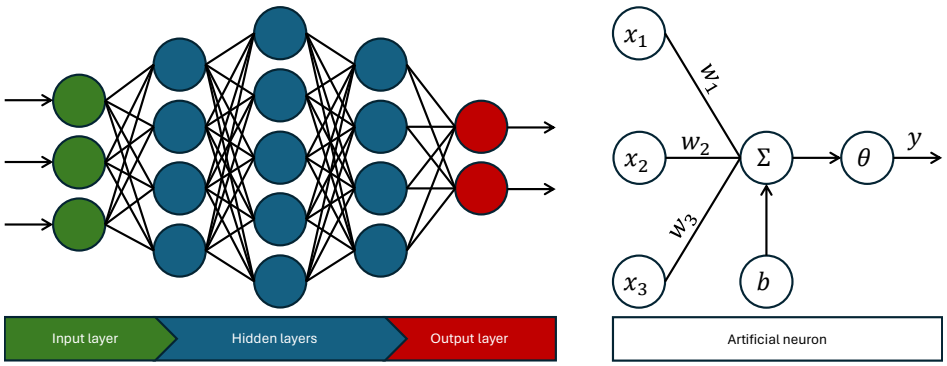


Figure 2.1: Depiction of a MLP. Artificial neurons, represented as circles, transform and propagate input values toward the output layers. A cryptic representation of an artificial neuron is also depicted.

for drawing conclusions. To facilitate learning, the algorithm iteratively processes datasets $\mathcal{D}$ through models to extract those insights and learn from mistakes. A model is a structure of rules, random variables and/or perceptrons. The most commonly used are support vector machines, random forest, decision trees, Bayesian networks, and neural networks [61]. Forward propagating the data points x through the model M serve for generating an output $\hat{y}$. Yet, in favor of learning, a recursion mechanism is established to reconfigure the internal variables of the model, also known as weights, back propagating the findings from the obtained output based on the calculated loss $\mathcal{L}$.

Neural Networks

Neural networks are a ML model type based on artificial neurons. An example of a neural network in shown in Fig. 2.1. Models are usually formed by concatenation of multi layer perceptrons (MLP), comprising a multitude of artificial neurons. These neurons are interconnected and transfer information from one layer to the next. In practice, a neuron receives a set of inputs x_i. Then, using a configurable set internal variables, referred to as weights w_i, the neuron applies a linear factor over the input. Neurons also use a bias term b to leverage its linear output, as well as an activation function θ to introduce non-linearity. Neurons within the same MLP layer are not directly connected; instead, the outputs from neurons

in the preceding layer serve as inputs for neurons in the current layer, and so forth. The output of a neuron can be expressed as:

$$y = \theta(\sum_i x_i w_i + b) \quad \forall i \quad \text{in the previous neuron} \tag{2.1}$$

2.1.2 Learning Techniques

There is a substantial amount of learning techniques in the repertory of machine learning. These differ on the utilisation of labels, or absence of them, to train models. The ones used in our work are presented in the following subsections.

Supervised Learning

Supervised learning involves the training of a model using labeled examples. During this process, input data points x and their corresponding labels y are made available for training and validation. Consequently, a dataset of N samples can be represented as a set $\mathcal{D} = \{(x_i, y_i), i \in [1, N]\}$.

Weakly-Supervised Learning

Weakly-supervised learning model involves training samples defined as a collection of instance data points with unknown labels, referred as bag. Although the true class of the instances is not given, the label of the group altogether is provided. A bag of instances $\mathcal{X} = \{x_l, l \in [1, L]\}$ is paired to a label y. Thus, the dataset is defined as $\mathcal{D} = \{(\mathcal{X}_i, y_i), i \in [1, N]\}$.

Active Learning

Active learning is driven by acknowledging that labeled examples do not hold equal importance.The practice prioritizes data that provides the most useful information to maximize its impact on training a supervised model. The most informative data is designated by means of a query strategy, often based on model classification uncertainty. Then, using an acquisition function, data points are sampled either through selection by an expert oracle or automatically chosen based on higher scores derived from the selected query strategy. Therefore, active learning comprises splitting a training set $\mathcal{X}_\mathcal{T} = \{x_i, i \in [1, N_T]\}$ into an initial small subset of the training

set $\mathcal{X}_l = \{x_j, j \in [1, N_0]\}$ and a pool of unlabeled data $\mathcal{P} = \{x_k, k \in [1, N_\mathcal{P}]\}$. Initially, the model is trained with the starting train set $\mathcal{X}_l$ and samples are adhered progressively from the pools $\mathcal{P}$. The training procedure relies on training the same model several times until a criteria is met. Deciding when to stop training involves balancing performance improvement with resource utilization. Common approaches include monitoring performance plateau, allocating labeling or computational budgets, assessing annotation efficiency or setting convergence criteria. Ultimately, the decision should be guided by the objectives, available resources, and limitations of the project.

Contrastive Learning

Contrastive learning has emerged as a promising technique for acquiring feature representations from extensive unlabeled data. In contrast to previous learning techniques that focus on classification, contrastive learning aims to train a feature extractor. Therefore, the focus relies on the loss function for feature representation comparison. Given a batch of N_b random samples from a training set of images $\mathcal{X}_\mathcal{T}$, a set of two image transformations are applied to all images in the batch in order to obtain $2N_b$ augmentations. These augmentations are forwarded through the model to obtain feature representations $\mathbf{z}$ and later compared using a contrastive loss function $\mathcal{L}_c$.

2.1.3 Loss Functions

This section highlights a comprehensive summary of loss functions used in the frame of this thesis.

Cross Entropy Loss

Cross entropy measures the distribution of the true label y and the predicted value $\hat{y}$. The binary cross entropy loss is defined as:

$$\mathcal{L}_{ce} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{2.2}$$

Focal Tversky Loss

The Tversky index (TI) leverages false predictions, emphasizing on recall in case of large class imbalance tuning parameters α and β. TI is defined as:

$$\mathrm{TI}_c(\hat{y}, y) = \frac{1 + \sum_{i=1}^{N} \hat{y}_{i,c} y_{i,c} + \epsilon}{1 + \sum_{i=1}^{N} \hat{y}_{i,c} y_{i,c} + \alpha \sum_{i=1}^{N} \hat{y}_{i,\hat{c}} y_{i,c} + \beta \sum_{i=1}^{N} \hat{y}_{i,c} y_{i,\hat{c}} + \epsilon} \tag{2.3}$$

where $\hat{y}_{i,\hat{c}} = 1 - \hat{y}_{i,c}$ and $y_{i,\hat{c}} = 1 - y_{i,c}$ are the probability that sample i is not of class $c \in \mathcal{C}$. ϵ is used for numerical stability, preventing zero division operations. Focal Tversky Loss (FTL) employs another parameter γ for leveraging training examples hardship:

$$\mathrm{FTL}_c(\hat{y}, y) = \sum_{c} (1 - \mathrm{TI}_c(\hat{y}, y))^{1/\gamma} \tag{2.4}$$

Contrastive Loss

Let $\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}_i^\top \mathbf{z}_j / \|\mathbf{z}_i\| \|\mathbf{z}_j\|$ be the cosine similarity between l_2 normalized vectors $\mathbf{z}_i$ and $\mathbf{z}_j$, then the loss function is defined as:

$$\mathcal{L}_c = -\frac{1}{2N} \sum_{i \in I} \log \frac{\exp(\mathrm{sim}(\mathbf{z}_\mathrm{i}, \mathbf{z}_\mathrm{j})/\tau)}{\sum_{k=1}^{2N} 1_{[k \neq i]} \exp(\mathrm{sim}(\mathbf{z}_\mathrm{i}, \mathbf{z}_\mathrm{k})/\tau)} \tag{2.5}$$

where $1_{[k \neq i]}$ is a binary indicator to indicate all other instances than i, and τ is a temperature coefficient to control the strenght of penalties on hard negative samples. $\mathcal{L}_c$ corresponds to the unsupervised version, although the supervised contrastive learning loss $\mathcal{L}_{sc}$ does exist. The supervised loss $\mathcal{L}_{sc}$ considers images from the same class in the batch, using the corresponding image label y, and does not punish the model for generating similar representations among images from the same class:

$$\mathcal{L}_{sc} = \sum_{i \in I} \frac{-1}{|P_{(i)}|} \sum_{p \in P_{(i)}} \log \frac{\exp(\mathrm{sim}(\mathbf{z}_\mathrm{i}, \mathbf{z}_\mathrm{j})/\tau)}{\sum_{k=1}^{2N} 1_{[k \neq i]} \exp(\mathrm{sim}(\mathbf{z}_\mathrm{i}, \mathbf{z}_\mathrm{k})/\tau)} \tag{2.6}$$

where $P_{(i)} \equiv p : y_p = y_i$, while $|P_{(i)}|$ represents its cardinality.

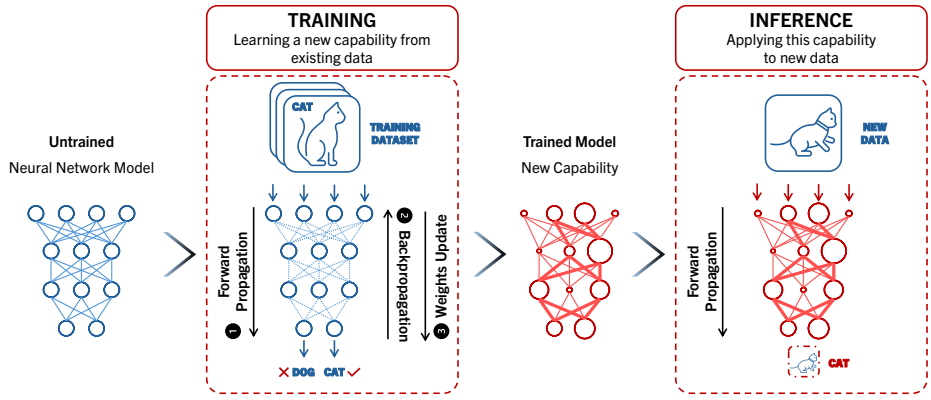


Figure 2.2: The neural network model undergoes a lifecycle, commencing with its untrained state. Through iterative training, the model acquires new capabilities by extracting insights from a designated training set. This iterative process involves feeding data forward through the network, utilizing backpropagation to adjust weights based on computed loss. Upon completion of the training phase, the model transitions to the inference stage, where it leverages its acquired knowledge to generate predictions on new unseen data.

From Training to Inference

The structure and training process of neural networks are pivotal for their functionality. However, it is crucial to highlight a notable distinction within their lifecycle: the transition from training to inference, as illustrated in Fig. 2.2.

The journey of a neural network begins with its inception, often in an untrained state. Through training, the network undergoes a transformative process wherein it learns to discern patterns and relationships within data, essentially acquiring the ability to perform specific tasks. This training phase is characterized by iterative adjustments to the network's weights, guided by an optimization process like backpropagation, gradually enhancing its performance.

Once the training phase concludes and the network achieves a satisfactory level of performance, it transitions into the inference stage. Here, the network's learned knowledge is put to practical use. When presented with new, unseen data, the network applies its acquired understanding to make predictions or decisions. Unlike training, where the emphasis lies on adjusting parameters to minimize error, inference focuses on leveraging the network's learned representations to generate accurate predictions.

2.1.4 Deep Learning

Deep learning (DL) is a sub-field of ML that utilizes deeper neural networks. The presence of additional layers permits more refined understanding of intricate concepts, while handling data with little to no preprocessing. With conventional neural network models, neurons can only establish connections with individual data points, leading to a significant number of connections (e.g., all pixel values in an image need to be connected with all input neurons). To mitigate this processing overload, machine learning models depend on meticulously crafted feature extraction strategies before initiating model training. In contrast, DL models do not require hand-crafted feature extraction [62]. In exchange for these advantages, DL algorithms require larger amounts of data to judge which features are important. They cover a wide range of applications, in terms of data modalities and disciplines. These include activity recognition [63, 64], natural language processing [65, 66], object segmentation [67, 68], signal processing [69, 70], privacy [71, 72], and many others.

Convolutional Neural Networks

Since DL models need to carefully extract features from the raw data, dedicated tools are utilized to accomplish this objective. In the case of images and video, it is common to use a convolutional neural network (CNN). CNNs are well-suited for the image modality, as convolutions are translation invariant and focused on local features. Also, CNNs contain lesser number of trainable parameters than a MLP would have given the vast number of pixels an image contains. Moreover, extensive research on image processing has led to a substantial number of pre-trained models to transfer learning from [73, 74]. A CNN is often formed of a series of convolutional and pooling layers. A convolutional layer conducts discrete convolutions on the input I using a learnable kernel $\mathcal{K}$. Here, $I(x_h, y_v) \in \mathbb{R}^2$ and $\mathcal{K}(k_1, k_2) \in \mathbb{R}^2$ represent a 2D image and a 2D kernel, with (x_h, y_v) and (k_1, k_2) denoting their width and height. The convolution operation between I and $\mathcal{K}$ is defined as:

$$g''(x_h, y_v) = \mathcal{K}(k_1, k_2) \circledast I(x_h, y_v) = \sum_{i=0}^{k_1-1} \sum_{j=0}^{k_2-1} \mathcal{K}(i, j) I(x_h + \widetilde{k_1} - i, y_v + \widetilde{k_2} - j) \tag{2.7}$$

where $\circledast$ denotes the convolution operation, and $\widetilde{k_1} \equiv \lfloor \frac{1}{2}(k_1 - 1) \rfloor$ and $\widetilde{k_2} \equiv \lfloor \frac{1}{2}(k_2 - 1) \rfloor$ correspond to the half-width and half-height of the kernel $\mathcal{K}$. The convolutional layer utilizes kernels and strides to generate feature maps capturing local patterns in the input data. Pooling layers then downsample feature maps spatially, reducing dimensionality and enhancing model robustness. Average-pooling and max-pooling are common pooling operations, calculating the average and maximum values from input regions to create downsampled feature maps, respectively. Ultimately, a CNN $G_w : \mathcal{I} \to \mathcal{H}$ with trainable weights θ is transforming input images into feature embeddings.

Multi-scale Models

Multi-scale models represent a versatile approach to data analysis, operating across multiple resolutions to capture hierarchical information and extract discriminative features. By considering data at various scales, these models enhance understanding by discerning fine-grained details alongside broader patterns [26, 31, 75, 76]. Additionally, multi-scale models offer adaptive analysis capabilities, allowing them to adjust the level of analysis based on task requirements or data characteristics. Overall, multi-scale models provide a flexible framework for analyzing complex systems, offering improved insights and performance across diverse domains. In this thesis, models that operate on a single scale are denoted as MONO, those on two scales as DI, on three scales as TRI, and so forth. An example of a multi-scale CNN input is represented in Fig. 2.3.

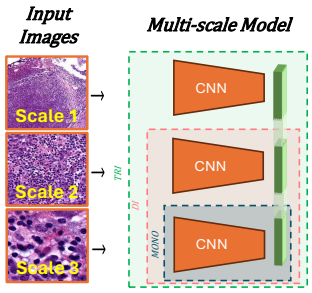


Figure 2.3: Illustration of a multi-scale CNN input, showcasing the convention of denoting models operating on varying scales. MONO refers to single-scale models, DI to dual-scale models, and TRI to triple-scale models.

Sequential and Attention Models

In regards to language processing, an architecture that supports the sequential nature of words needs to be adopted. Typically, recurrent neural networks (RNNs) and long short-term memory (LSTM) architectures are utilized [77]. These architectures thrive in sequential data-handling and understanding the contextual information. An evolution of these sequential networks emerged in transformers [78]. The effectiveness of transformers is rooted in their capacity to capture long-range dependencies, process data in parallel, and leverage the self-attention mechanism. Transformers quickly have become widely popular, and adaptations have crossed disciplines for imaging [79, 80] and video [81]. However, transformers demand even larger amounts of data to achieve satisfying performance. In the medical domain, severe data limitations hinder its widespread adoption [82]. It's noteworthy that the use of attention mechanisms for image processing, the main appeal of the model architecture, predates the incorporation in vision transformers [83]. It was found that attention-based learning increases predictive performance and enhances feature discrimination in histopathological applications [84]. Moreover, attention scores influence forward propagation. An attention score a_i for a feature embedding $\mathbf{h}_i$ can be calculated as:

$$a_i = \frac{\exp\{\mathbf{w}^\top(\tanh(\mathbf{V}\mathbf{h_i}^\top) \odot \mathrm{sigm}(\mathbf{U}\mathbf{h_i}^\top))\}}{\sum_{l=1}^{L}\exp\{\mathbf{w}^\top(\tanh(\mathbf{V}\mathbf{h_l}^\top) \odot \mathrm{sigm}(\mathbf{U}\mathbf{h_l}^\top))\}} \tag{2.8}$$

where L is the number of elements in a set of data points, $\mathbf{w} \in \mathbb{R}^{L\times 1}$, $\mathbf{V} \in \mathbb{R}^{L\times M}$ and $\mathbf{U} \in \mathbb{R}^{L\times M}$ are trainable parameters and $\odot$ is an element-wise multiplication. Furthermore, the hyperbolic tangent $\tanh(\cdot)$ and sigmoid $\mathrm{sigm}(\cdot)$ are included to introduce non-linearity for learning complex applications. The main advantages of attention in imaging relies on the flexibility to decide which data points are highly informative, and the interpretability provided by the output attention scores. Furthermore, during the inference stage, we can visualize these scores to gain deeper insights into the model's reasoning, thereby enhancing its interoperability and addressing the challenges posed by the "black box" problem.

Part II

Clinical Application

Chapter 3

Medical Application

In this chapter, we introduce bladder cancer disease, diagnostic assessment, risk factors for analysis and treatment strategies.

3.1 Histopathological Workflow

Figure 3.1: Histopathological workflow. Tumour samples are extracted from a biopsy, followed by a histological process to stabilize the biological tissue. The tissue sample is sectioned, stained, and later, digitally scanned, for visual examination. Analysis of image features and clinicopathological data leads to a report that dictates disease management. Image reprinted with permission from QPS Neuropharmacology [85].

A pathologist examines the tissue to provide a diagnosis that will result in a treatment intervention for a given patient. Understanding the histopathological process is beneficial for more accurate diagnostics and prognostics, as it provides insights into the underlying cellular and tissue-level changes associated with various diseases, enabling clinicians to make informed and precise decisions. This same understanding serves as foundational knowledge for guiding the development of computational pathology algorithms.

The clinical histology process begins when a treating physician determines histological confirmation necessary in order to evaluate and determine the

appropriate treatment procedure, or a when a tissue sample is collected. Tissue samples are extracted from the patient, usually via a biopsy or excision, so a pathologist can evaluate in detail the biological properties of the sample [86]. After a tumor tissue is extracted, the tissue is chemically and physically stabilized. The tissue is first immersed into a fixative solution that will stop cellular degradation and prevent microorganism proliferation. After fixation, the tissue is dehydrated and cleared for paraffin infiltration. The paraffin is liquefied to infiltrate the tissue and later cooled down to keep the tissue firm. A block of paraffin is formed after pouring the liquid over the tissue that is placed on the bottom of a mold and left onto a cooling plate to harden the paraffin. Positioning of the tissue during this step is fundamental for sectioning in the correct orientation. Proper orientation helps on visualizing the resected area that otherwise wouldn't have been visible. A microtome cuts thin slices of the tissue paraffin block. These sections are plunged onto water slightly about paraffin melting point to prevent tissue wrinkling and expand it. Once the section is heated, it is mounted on glass slides for later examination. Slices are normally 3-4μm thick to display relevant cellular local information, since thinner sections are harder to extract and usually are cauterized and thicker slides tend to make staining dark and lose contrast which makes nuclear detail observation difficult.

After undergoing this process, the tissue presents little contrast and biological diversity is almost invisible for the naked eye. Tissue samples are stained to create contrast among different tissue types and cellular elements. Most staining procedures are based on dyes that will highlight certain cellular features. The amount of dye used should be interdependent of the thickness of the sectioned tissue. For very thin slices, lower amounts of dye are recommend to not saturate the coloring, while a minimal doses of dye would result in low contrast and hardly visible differentiation between cellular components. On clinical practice, the chemical properties of these dyes produce a desired visual appearance based on the components to be evaluated. The most commonly used stain for diagnostic purposes is hematoxylin and eosin (HE) stain [87, 88]. HE provides excellent contrast between cellular, for which hematoxylin turns cell nuclei purple wherein eosin turns cytoplasm pink. Some studies have used saffron additionally (HES) for further enhancing the visual characteristics of morphological structures of the lamina propria, mainly for contrasting pink cytoplasm [89–91]. Labs usually section multiple adjacent slices from the same sample to observe different patterns and morphological features using different stains,

also called immunohistochemistry. Immunohistochemistry is a procedure that aims to detect, amplify and make visible a specific antigen, which is usually a protein. This technique makes it possible to identify the localization of a specific substance at the tissue or cellular level, based on the use of antibodies that specifically bind to a substance to be identified. Emphasis is also given to laboratory applications with the aim of providing a valuable support to medical diagnostics, which in the case of histopathology is to underline the presence of immune cells within the tissue, being the latter a positive prognostic indicator.

Traditionally glass slides are stored in a mayor archive for later selection and observation through a microscope. Nowadays, glass slides are scanned and digitized into whole slide images (WSIs), which allows pathologists to examine tissue through computer screens. A pathologist analyzes possible abnormalities that may be present including tissue architecture, texture and presence of malignant cellular types, among others, to elaborate an accurate diagnosis. Based on the evidence gathered from the histological images, and molecular data from sequencing techniques, the pathologists elaborate a study report on the patient current status and follow-up procedures for eliminating the disease.

3.2 Bladder Cancer

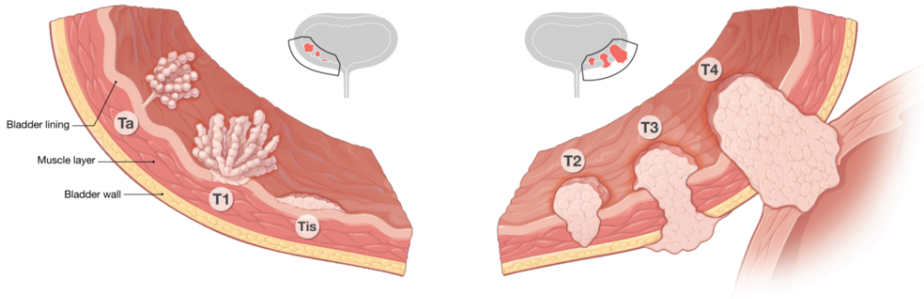


Figure 3.2: Stages of bladder cancer. At the left-hand side, we have non-muscle invasive bladder cancer (NMIBC), for which we have flat lesions (Tis), and papillary lesions (Ta, T1). Ta is a non-invasive carcinoma while T1 invades the subepithelial connective tissue. On the right-hand side, we have muscle invasive bladder cancer (MIBC). At T2 stage the tumour has invaded the muscle layer, T3 the perivesical tissue and T4 adjacent organs. Image reprinted with permission from Roswell Park - Comprehensive Cancer Center [92].

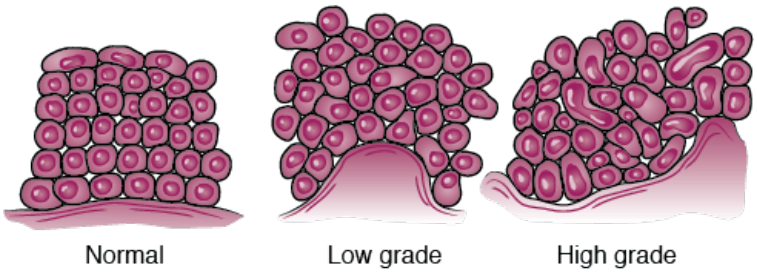


Figure 3.3: Grading of bladder cancer. At the left-hand side, normal or healthy urothelium. The middle and right-hand side cellular structures represent low- and high-grade urothelial tissue, according to the WHO04 grading system. With increasing grade, the cellular arrangement becomes more intricate, leading to a greater degree of cell entanglement and increased amorphousness. Image reprinted with permission from Bladder Cancer Academy Network [93].

Bladder cancer (BC) is a heterogeneous disease emerging from the urothelial lining of the bladder. In Fig. 3.2, we highlight examples of tumours at different T-stages, following the Tumour Node Metastasis (TNM) staging system. TNM defines the stage of the cancer according to the deepest layer invasion, ranging from tumours contained within the urothelial lining (Tis) to adjacent organ invasion (T4) [94]. As the stage increases, the prognosis becomes more unfavorable. BC is further subdivided into two major subcategories based on muscle invasion: non-muscle invasive bladder cancer (NMIBC), encompassing Ta, T1 and Tis tumours; and muscle invasive bladder cancer (MIBC), including T2, T3 and T4 tumours. NMIBC prognosis, hence treatment options, are vastly more positive than MIBC [95]. Patients diagnosed with MIBC are urgently prompted with radical cystectomy surgery, meaning that the entire bladder and contiguous lymph nodes are removed, resulting in increased morbidity and reduced quality of life.

Grading refers to the degree of differentiation of cancer cells in comparison to non-cancerous cells [96]. Nowadays, the World Health Organization has two grading systems in place. WHO73 proposes three grades, namely 1, 2 and 3; while WHO04 proposes low- and high grade. Recent studies show a benefit in using WHO04 over WHO73, although hybrid classification systems have been prevalent [97, 98]. For this reason, in this thesis, we will proceed with WHO04. An example showing low- and high grade urothelial tissue in comparison to normal urothelium is shown in Fig. 3.3. Nevertheless, a significant component of inter- and intra-observer variability

exists for either of the classification systems [99, 100]. All MIBC cases are high-grade, thus grading carries no prognostic value [36].

In this thesis, we are solely focusing on NMIBC, as 75% of newly diagnosed patients with BC are NMIBC [101]. Moreover, NMIBC patients are eligible for treatment that may effectively eliminate any remaining cancer cells left in the bladder. Active surveillance is advised after the first transurethral resection of a bladder tumour (TURBT) for potential recurrence or progression of the disease [102, 103].

3.2.1 Diagnosis and Risk Factors

Haematuria, which refers to the presence of blood in the urine, is often the earliest indication of NMIBC disease [104]. Several non-invasive imaging techniques are adopted for determining the presence of NMIBC, being computed tomography [105], ultrasound [106] and magnetic resonance imaging [107]. However, histopathological analysis remains the ultimate test for disease risk assessment, hence it remains as the gold standard due to its precise diagnostic capabilities [108].

The main diagnostic risk factors for BC occurrence include external and genetics [109]. Smoking tobacco is the leading risk factor, followed by exposure to gases rich in carcinogenic chemical compounds and chlorinated liquids. The impact of diets is limited, yet highly saturated fats are linked to an increased incidence rate [110].

3.2.2 Prognosis

In pursuance of predicting the risk of disease recurrence and progression, several prognostic models have emerged. These models rely on a scoring system with significant clinical and pathological factors. The most prominent factors are age, gender, number of tumours, grade, concomitant CIS, tumour size and prior recurrence. The most used scoring models are introduced by the European Organisation for Research and Treatment of Cancer (EORTC), Club Urológico Español de Tratamiento Oncológico (CUETO) and European Association of Urology (EAU) [10, 111, 112].

Based on the 2022 EAU NMIBC scoring model, patients are categorized into risk groups according to their likelihood to progress to MIBC. The calculation is formulated based on an analysis of individual patient data, considering age>70, multiple papillary tumours and tumour size >3cm

as additional clinical risk factors. This risk stratification is essential for a tailored treatment and intervention. Finally, DNA and RNA genetic alterations result in molecular subtypes with distinct prognostic implications [113].

Risk groups prescribe guidelines for appropriate disease management. Low-risk patients benefit from a single instillation of chemotherapy for reducing the recurrence rate, while for intermediate-risk patients repeated instillations are required for recurrence free survival [114]. In addition, several studies agree that intravesical Bacillus Calmette-Guérin (BCG) immunotherapy results in more positive outcomes for intermediate, high and highest-risk patients [115–117] For intermediate-risk disease, guidelines typically recommend a 1-year course of BCG treatment. This treatment protocol typically begins with an induction phase comprising six weekly instillations, followed by a maintenance phase involving additional regular instillations over a predetermined period. For patients with high-risk of disease progression 1-to-3 years of BCG treatment is indicated, although radical cystectomy is suggested. Immediate radical cystectomy for highest-risk is advised, albeit the patient may opt for regular BCG therapy. Nonetheless, NMIBC may exhibit inadequate treatment response or experience recurrence despite an initial positive response [11]. In those cases, radical cystectomy is the established procedure of choice.

3.3 Computational Pathology

As slide scanning technology becomes faster and more reliable, the availability of WSI data for training convolutional neural network models is rising. Integrating clinical information, biomarkers, and sequencing data, computational pathology (CPATH) is set to become standard practice [118]. This approach streamlines pathology workflows and provides a comprehensive view, aiding pathologists in monitoring complex disease progression for better patient care [119].

3.3.1 Whole Slide Images

Whole slide images (WSIs) are high-resolution digital files from a scanned microscope slide of sectioned tissue biopsy. The nature of these images is of voluminous size, as a typical size of 200 000 $\times$ 200 000 can be found for 800x magnification, as the visual appearance of the tissue has been

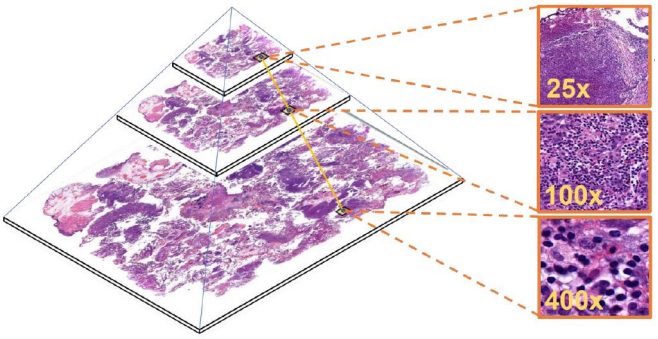


Figure 3.4: Whole-slide images (WSIs) are stored in a pyramidal format, with the highest magnification level represented by the base image. Here, a set of three tiles with the same center physical point are extracted at different magnification levels.

enlarged 800 times. For this reason, WSIs are also referred as gigapixel images [8, 120–122].

WSIs are stored in a pyramidal structure, where the original glass slide is scanned at the highest resolution, and multiple down-sampled versions are retained. This emulates the behavior of a traditional microscope for pathologists, facilitating rapid zooming in and out. We refer to magnification as the ratio between the apparent size of the tissue and its actual size. The usual base magnification used is 400x, although others are also adopted [44, 45, 123]. It is worth noting some might use either 40x or 400x to describe the same level of magnification. The roots of this discrepancy originate from historical microscope conventions. Traditional microscopes labeled objectives relative to the combined magnification with a 10x eyepiece, resulting in a 40x objective being labeled as providing 400x total magnification [124]. In digital imaging, however, the terminology may refer solely to objective magnification, leading to different labels for the same level of magnification. Moving forward, we will use the total magnification. A visual representation of the pyramidal nature of WSIs is shown in 3.4.

Due to the bare size of the WSIs, tiling is a essential step for their efficient analysis in CPATH tasks. A sliding window algorithm is adopted for extracting tiles from the raw slides. These tiles are individually processed, which enables processing for computing hardware while reducing computationally intensive and memory-consuming operations. Also, selectively extracting tiles from selected regions diminishes the need to load the entire slide, permitting the exclusion of non-diagnostically relevant tissue areas and magnification selection.

3.3.2 Image Analysis and Applications

Extensive work on assessment of slide quality and pre-processing has been a main point of interest in the research community [125–127]. As the surgical procedures for tumor removal are intrinsically complicated, several artifacts and other undesired tissue come to be present at WSIs [128]. Works on artifact removal focus on the detection and removal of blur, damaged tissue and tissue folds from the slides [129–133], although others focus on the overall evaluation of the slide quality for discarding artifact noisy WSIs [134–137]. Nonetheless, artifact-free WSIs still suffer from stain variability. Approaches to mitigate the color variability include augmentation [138, 139], normalization [140–143], and deconvolution to obtain separate images for each stain [144–146]. Another emerging trend in this regard is the use of generative deep learning models for potentially addressing the aforementioned challenges [147, 148]. A software tool for quality control of WSIs is HistoQC [149]. It automatically detects and flags common issues such as artifacts and stain abnormalities, resulting in a valuable tool for ensuring the reliability of slides.

An overview of software tools for digital pathology is provided by Guerrero et al. [150]. Among the most prominent tools we find Ilastik [151], ImageJ [152], and CellProfiler [153]. Ilastik provides capabilities for image classification and segmentation, employing an intuitive machine learning toolbox with support for both pixel-level and patch-level analysis. ImageJ is a popular open-source software tool used in medical image analysis. CellProfiler, on the other hand, is a versatile tool with a user-friendly interface for quantifying cell features. Another tool is QuPath, a digital image analysis software platform with integrated machine learning algorithms that can be trained by users. [154]. Additionally, Visiopharm, a prominent company in the field of digital pathology, offers a range of software solutions for image analysis, quantification, and interpretation. Notable works employing Visiopharm's software include studies focusing on image-based biomarker assessment, tissue morphometry, and digital pathology workflow optimization [155–158].

Regarding CPATH algorithms emphasizing on bladder cancer, we observe a significant lack of research in comparison to the other more popular cancer types, like breast and prostate, among others [18, 19, 23, 159, 160]. In the realm of bladder cancer diagnostics, numerous applications have been explored, each contributing to different aspects of histological image analysis. Notably, tissue segmentation has garnered attention. Wetteland

et al. [26, 27] employed a multi-scale CNN input for segmenting all tissue into five distinct classes using patches at different magnification levels; while Niazi et al. [28] used a U-Net for pixel-level segmentation of distinct pathology elements. Similarly, advancements in cell segmentation have been made, as evidenced by the works of Zheng et al. [29], who employed QuPath in-built functions to preprocess the slides and extract cell features for training a naive neural network. Schmidt et al. [30] proposed a method based on star-convex polygons, a more suitable method for detecting round shapes of cell nuclei. Results demonstrate that the proposed solution outperformed state-of-the-art pixel-segmentation deep learning methods like U-Net and Mask R-CNN. In the context of grading, another critical aspect of histological analysis, significant progress has been achieved. Wetteland et al. [31] use a multi-magnification CNN backbone for grading NMIBC urothelium patches. Then, it leverages individual patch predictions into a WSI-level prediction using a voting mechanism. Barrios et al. [32] employ an ensemble of four weight-independent CNNs to predict grade and stage, to ultimately correlate the predictions into risk hazard. Zhang et al. [33] introduce an automated pipeline employing deep learning models. This pipeline delineates lesion areas, generates reports based on the identified lesions, and ultimately provides a grade prediction at the WSI-level. In regards to staging methods for bladder cancer, only one work was found. Yin et al. [34], from pathology-annotated ROIs, used CellProfiler for segmenting cells and ImageJ for extracting features from those cells. Subsequently, various neural networks were employed to ascertain the stage of the ROI, with probabilistic models demonstrating superior performance.

In contrast to diagnostics, prognostic methods have not received as much research attention. However, several crucial applications aim to discern clinical outcomes based on histological images. Recurrence prediction has been a focal point for researchers. In Tokuyama et al. [37], pathologists manually extracted ROIs from relevant WSIs, to then use Ilastik to segment cell nuclei and CellProfiler to measure nuclear morphologic and texture features. Ultimately, extracted nuclei features are fed through support vector machines and random forest algorithms to predict the event of tumor recurrence. In Urdal et al. [38], a pathologist similarly delineates a ROI per WSI. Features are subsequently extracted from the histogram and inputted into a classifier. In Lucas et al. [161], they exploit a pre-trained CNN for feature extraction of lesion areas and a bidirectional gated recurrent unit for predicting recurrence. They also delve into the impact of clinicopathological data, and they found that an hybrid approach of

clinical and histological features worked best. Unfortunately, progression prediction has not been directly addressed by means of histopathological image analysis. Nevertheless, survival prediction studies, such as those conducted by Chen et al. [40, 41], have aimed to provide insights into long-term patient outcomes. They used QuPath to segment cell nuclei and CellProfiler to extract nuclei features, to subsequently use a machine learning model for risk stratification. Additionally, the assessment of response to treatment has been investigated by Mi et al. [42], where they used a mixture of QuPath and the previously discussed work in [30], to delineate nuclei and extract features, to consequently employ clustering methods for grouping distinct responses. This work further contributed to the prognostic modeling in bladder cancer.

Chapter 4

Data Material

This chapter describes the data material used in this work. The majority of the data used consists of histological whole slide images, some of which are annotated by pathologists, along with associated clinicopathological data from the corresponding patients. The image dataset overview for each of our contributions is also presented.

4.1 Dataset Overview

The overview of the image data material is described in the following section. In this thesis, we have used public and private datasets. We refer to the private medical datasets using capital letters for increased readability. This notation is extended to the rest of the thesis. The private datasets were provided by either Erasmus Medical Center (EMC), Rotterdam, The Netherlands, or Stavanger University Hospital (SUH), Stavanger, Norway.

For training and evaluating the different models, multiple training, validation, and test sets were created. While some WSIs appear in multiple datasets, precautions were taken in all experiments to avoid cross-contamination between training and testing sets of the same model.

4.2 Public Datasets

4.2.1 MNIST

The MNIST (Modified National Institute of Standards and Technology) database is a classic image dataset consisting of handwritten digits [162].

MNIST consists of tiles of size 28×28, with a training and test set of 60,000 and 10,000 examples, respectively. The dataset contains 10 classes corresponding to the 10 existing digits represented in the images.

In Paper I, the models were trained and tested with 20,000 and 5,000 bag-of-bags, respectively, constructed by randomly extracting 16 instances within the class of desired instance labels. The digit 9 was prompted as the positive instance class.

4.2.2 PCAM

PCAM (PatchCamelyon) is an image dataset consisting of patches extracted from WSIs of lymph node sections [163, 164]. PCAM consists of 327,680 patches of size $96 \times 96 \times 3$, where each patch is annotated in a binary manner to indicate the presence of metastatic tissue.

In Paper I, the models were trained and tested with 20,000 and 5,000 bag-of-bags, respectively, constructed by randomly extracting 16 instances within the class of desired instance labels. The presence of metastasis was prompted as the positive instance class.

4.3 Private Cohorts

The database from EMC is a multi-center cohort of HE stained WSIs of HR-NMIBC. A total of 779 WSIs were scanned using a 3DHistech P1000 scanner at 800x magnification stored as MRXS files. The cohort from SUS is a retrospective study of NMIBC patients. The data consists of 314 WSIs, with either HE or HES staining, digitized using a Leica SCN400 scanner at 400x magnification stored in the SCN file format.

In computational pathology and medical imaging modalities, labels can be region-based (e.g. identifying tissue types within images) or patient-based (reflecting disease progression or treatment response). Some labels can serve both purposes (e.g., tumor grade), while others are inherently region- or patient-based. To integrate both label types, we adopt annotations over WSIs for region-based labels and clinicopathological data for patient-based labels.

Figure 4.1: Microdraw is a web-based tool for annotation of large histopathological images. The image shows the different classes considered in our annotation guidelines as well as an illustration of annotated areas by a pathologist.

4.3.1 Annotations

Labeled ground truth holds major significance for training models, but more so for validating them. Unsupervised and weakly supervised strategies offer learning opportunities without detailed annotations, but a test set annotation can be crucial for model evaluation. In order to incorporate annotations, along with our project collaborators at Univesitat Politècnica de València (UPV), we implemented a web-tool for histological image annotation named Microdraw, based on the OpenSeadragon library [165]. The main goal of using a web-based annotation software was to preserve our WSIs locally, but remotely accessible through the portal. The remote annotation reduces data transfers and quickens the overall process. This tool

allows free-hand drawing as well as point insertion for defining an enclosed area. Different tissue type and subclass labels can be assigned to the drawn areas (e.g., 'urothelium: high grade', 'lamina propria: with tumour infiltrating lymphocytes'). The annotation guidelines were structured to enable untrained pathologists to conduct the majority of annotations in the NMIBC dataset. Specifically, a non-expert would undergo training to classify tissue types (e.g., urothelium, lamina propria, muscle, blood and artifacts), after which an expert pathologist would further categorize subclasses and/or grade a portion of the dataset.

4.3.2 Clinicopathological Data

All WSIs collected were accompanied by a comprehensive set of clinical labels ranging from grading and staging, to follow-up data on recurrence and disease progression. However, only slides collected from the EMC cohort contained a wider set of patient demographics and treatment strategies.

It is imperative to delineate between distinct categories of clinical labels. These categories encompass labels employed directly for the ultimate classification of WSIs and those utilized to provide supplementary information, complementing the analysis of image features. In this thesis, the WSI-level labels included WHO04 grade during the first TURBT, recurrence with a median follow-up of 82 months, and BCG treatment response for up to 3-years follow-up. As for the complementary labels used, we included age, gender, smoking status, grade WHO04, grade WHO73, stage, tumor focality and size of the tumor. Note that grade was listed under both categories, as it can either serve as the primary objective for analysis or offer supplementary information for prognostic purposes.

4.3.3 Private Datasets

A comprehensive overview of the datasets created from the private cohort data is presented. Table 4.1 summarizes the dataset information.

Dataset A

Dataset A contains 155 WSIs from EMC, and it served for training and validating models presented in Paper II. A pathologist was asked to annotate areas in a rough, imprecise manner in order to obtain several annotated regions per WSI, within a time limit. The time usage was limited to a

Table 4.1: Overview of private cohorts. We outline the datasets and associated papers they were utilized in. The total number of WSIs is specified, along with their distribution across training, validation, and test sets respectively. Additionally, details regarding the region- and WSI-based labels employed are provided.

Cohort	Dataset	Papers	WSIs	Region labels	WSI labels
EMC	A	II	155 (127/16/12)	Blood Lamina propria Muscle Urothelium Artifacts	-
	B	III	37 (27/10/-)	Invasive cancerous area Tumour infiltrating lymphocytes Blood Lamina propria Muscle Urothelium Artifacts	-
	D	V	503 (399/52/52)	Invasive cancerous area Tumour infiltrating lymphocytes Lamina propria Urothelium WHO04 grade WHO73 grade	BCG treatment response
SUH	B	III	14 (10/4/-)	Invasive cancerous area Tumour infiltrating lymphocytes WHO04 grade WHO73 grade	-
	C	IV, V	300 (220/30/50)	WHO04 grade	Recurrence, WHO04 grade

maximum of one hour per WSI. As a result, some regions were annotated in every WSI, but not the entirety of present tissue was annotated. The aim of this annotation protocol was to collect diverse scenarios, hence capture the tissue heterogeneity characteristic from bladder cancer. In total, 127, 16, and 12 WSIs were annotated and used for training, validation and test, respectively.

Dataset B

Dataset B contains 37 WSIs from EMC and 14 from SUS, for staging in Paper III. The combined dataset included 23 non-invasive and 28 invasive tumours. All slides were partially annotated and revised by pathologists, with the annotations serving as ground truth. Only representative regions were annotated from each WSI; thus, no slide was fully annotated. Annotated regions from EMC included mainly tissue types (urothelium, lamina

propria, blood, muscle and artifacts), while SUH annotations contain primarily urothelium grading. In some cases, these tissue annotations might come with a subclass, such as grading of urothelium, presence of tumour infiltrating lymphocytes (TILs) in both urothelium and lamina propria, ICA, artifact type, among others.

Dataset C

Dataset C contains 300 WSIs from SUH for training and evaluating the grading models in Paper IV. All WSIs were rigorously graded in accordance with the WHO04 classification system, as either low-grade or high-grade, thus providing valuable slide-level diagnostic information. However, they lack location annotations pinpointing the precise areas of low- or high-grade regions within the WSI. As a result, the dataset is considered weakly labeled. The dataset employed in this study was divided into three subsets: 220 for training, 30 for validation, and 50 for test. In addition, pathologists annotated a set of 14 WSIs of NMIBC patients from the test set with low- and high-grade areas.

The WSIs used in Dataset C for train, validation and test, with the same distribution, are also used for recurrence prediction in Paper V. The prognostic labels correspond to recurrence or no recurrence. No WSIs were annotated, hence only weak labels regarding recurrence outcome were available.

Dataset D

Dataset D contains 503 WSIs from 453 patients, all from the EMC cohort, for prognostic models in Paper V. Since the treatment outcome is related to the patient as a whole and not to a specific region of the tumor, patches from various WSIs that belong to the same patient were merged as a single entity. A trainee pathologist annotated 217 WSIs, although none of the WSIs were fully annotated due to time and database storage constraints. Annotations contain tissue types, artifacts, grading and staging. Moreover, a detailed report of clinicopathological information per patient was disclosed with info about number of BCG instillations, responsiveness, among others.

4.3.4 Ethical Approval

For the EMC cohort, we refer to the ethical approval from Daily Board of the Medical Ethics Committee Erasmus MC, Rotterdam, The Netherlands, METC number: MEC-2019-704.

For the SUH cohort, we refer to the ethical approval from Regional Committees for Medical and Health Research Ethics (REC), Norway, ref.no.: 2011/1539, regulated in accordance to the Norwegian Health Research Act.

Part III

Contributions

Chapter 5

Learning Methods

In this chapter, we present the methods proposed in the outline of this thesis. We aim to define all the methods that serve as the foundation for all applications investigated. Paper 1 is dedicated to the topic, and the main methods, results, and contributions will be presented. In the thesis, Paper 1 is part of sub-objective SO_1 on utilizing non-supervised learning approaches to address the lack of annotations.

5.1 Active Learning

Labeled data is often scarce or costly for medical applications. An alternative to be more efficient on the usage of labeled data comes from using active learning. Learning and sample selection typically revolve around uncertainty metrics. We want to emphasize that missing a positive instance is detrimental, which hold true for medical diagnostics. The proposed active learning approach is explained in detail, see Algorithm 1. A training set $\mathcal{X}$ is divided into initial training subset $\mathcal{X}_0$ and pools of training data $\mathcal{P}$. The proposed active learning algorithm selects new samples based on model performance. Considering the false negative ratio (FNR) on the validation set $\mathcal{V}$ and total sample size S, N_j^i samples from the pools $\mathcal{P}$ are appended into the current training set, for every class i on iteration j. Furthermore, we define a relative class balance size δ_i^j to represent the percentage of data points for class i on the current training set $\mathcal{X}_j$. These steps are iteratively repeated until $\mathcal{P}$ cannot provide more samples of the requested class or J iterations have passed.

Algorithm 1 Active Learning Train Procedure

procedure $ALtrain(\mathcal{X}, \mathcal{V}, J, S)$
$\mathcal{X}_0, \mathcal{P} \leftarrow \mathcal{X}$ ▷ *Split train set into initial subset and pools*
while $j < J$ **or** $len(\mathcal{P}_i) > N_j^i$ **do**
  for $j \leftarrow 1$ to J **do**
    $MODEL_{TRAIN}(\mathcal{X}_j)$
    $MODEL_{EVAL}(\mathcal{V})$
    for $i \leftarrow 1$ to I **do**
      $N_j^i = \frac{FNR_{j-1}^i \cdot S}{\sum_i FNR_{j-1}^i}$ ▷ *Per class new data points*
      $\mathcal{X}_j^i \overset{+}{\leftarrow} sample(\mathcal{P}, N_j^i)$
    end for
  end for
end while
return $\hat{y}$

5.2 Weakly-Supervised Learning

In the medical imaging domain, labeled data is scarce or expensive to obtain. As a result, enabling the development of models in scenarios where fully supervised learning may be impractical or infeasible. Weakly supervised learning is valuable for training machine learning models with minimal labeled data, utilizing global image labels and disregarding annotations, cutting down on time and costs. Multiple instance learning (MIL) is a weakly supervised method trained in a supervised manner considering outermost bags and their corresponding labels. For the conventional MIL setting, a dataset $\mathcal{X}, \mathcal{Y} = \{(\mathbf{X}^i, y^i), \forall i = 1, ..., N\}$ is formed of pairs of sample sets $\mathbf{X}$ and their corresponding labels y, where i denotes the current sample for a total of N samples. A sample $\mathbf{X}$ consists of a bag of instances $\mathbf{x}_l$ of feature embeddings:

$$\mathbf{X} = \{\mathbf{x}_l, \forall l = 1, ..., L\} \tag{5.1}$$

where L is the number of instances in the bag. In order to obtain an instance representation or embedding from input data, for example an image patch, we make use of a feature extractor. An image feature extractor $G_f : \mathbf{i} \rightarrow \mathbf{x}$ is typically a convolutional neural network which maps an image patch $\mathbf{i}$ into a feature embedding vector $\mathbf{x}$. Instance representations $\mathbf{x}$ from a bag $\mathbf{X}$ are aggregated to form a bag representation using an aggregation

function Ξ, which can, for example, be replaced by either the *mean* or *max* operators.

A label $y \in \{0, 1\}$ is associated with the bag $\mathbf{X}$. Although each instance, $\mathbf{x}$, might be associated with a label y_l, they are generally unknown; hence only bag labels are used during training. At the inference stage, a test set might be associated with labels both at the bag and instance level to provide performance metrics. Under the conventional binary classification MIL assumption, a bag label is positive with the single presence of a positive instance. Then, the model learns comparing y with the prediction $\hat{y}$, computed by the bag classifier Θ_c.

$$y = \underset{l}{max}\{y_l\} \tag{5.2}$$

$$\hat{y} = \Theta_c(\Xi(\mathbf{X})) \tag{5.3}$$

Conventional weakly supervised models may overlook spatial context, which is intrinsically necessary for image processing. Therefore, more refined techniques can help address this issue by grouping extracted patches from regions separately during model training.

5.2.1 Paper 1 - Weakly Supervised Nested Model Architecture

In this section, the contributions in Paper 1 are presented. This paper presents a novel machine learning model architecture. Nested Multiple Instance with Attention (NMIA) overcomes the intricate relationship between instance and bag labels with nesting, while offering a high degree of interpretability using attention mechanisms. The datasets utilized correspond to Datasets MNIST and PCAM, as listed in Chapter 4.2.

The concept of bag-of-bags, or a nested setting, consists of levels of bags within bags where only the innermost bags contain instances. In Figure 5.1, the idea of grouping instances in inner-bags, inner-bags in larger bags, and finally in one outermost bag is illustrated. Let J denote the number of nested levels, K_j number of bags at level j and $L_{j,k}$ denote the number of instances or bags in a bag k. A set $\mathbf{X}_j$ contains a set of inner-bags $\mathbf{X}_{j,k}$:

$$\mathbf{X}_j = \{\mathbf{X}_{j,k}, \forall k = 1, ..., K_j\} \tag{5.4}$$

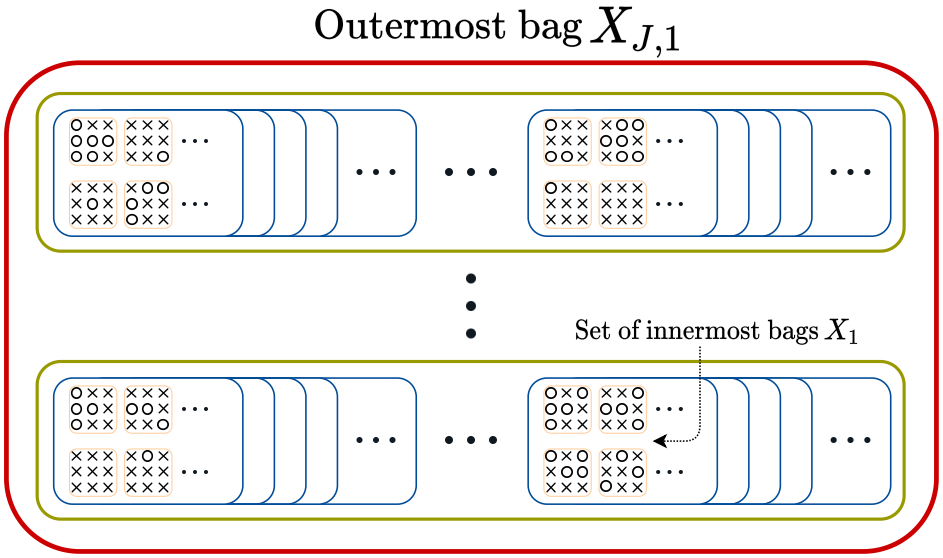


Figure 5.1: Nested bag-of-bags. Instances from two different classes (crosses and circles) are drawn into the innermost bags to form sets which are recursively grouped finally forming the outermost bag.

for j defining the current nesting level up to J levels of nesting. For $J = 1$, one level of nesting, the notation corresponds to the ordinary MIL notation described in Eq. (5.1). The number of inner-bags K_j can vary from level to level. For a given inner-bag k at level j, where l defines the instance number up to $L_{j,k}$, a bag of instances $\mathbf{X}_{j,k}$ is expressed as:

$$\mathbf{X}_{j,k} = \{\mathbf{x}_{j,k,l}, \forall l = 1, ..., L_{j,k}\} \tag{5.5}$$

By latent labels, we refer to the actual, but in general unknown, labels of an instance or inner-bag. When defining the dataset, however, these are necessary to determine the final bag-of-bags weak label y. For simplicity, we will refer to an instance latent label $y_{j,k,l}$ as y_l^j to describe the latent label of an instance l in a level j, omitting bag index k.

According to Ilse et al [84], a multiple instance attention block is constructed using an embedding-level approach to obtain a bag-level representation from the instances within, as depicted in Figure 5.2. The proposed attention module's input corresponds to low-dimensional embeddings $\mathbf{x}$. These are generated by a feature extractor G_f or by bag embedding representations of previous levels. The attention module analyzes those embeddings

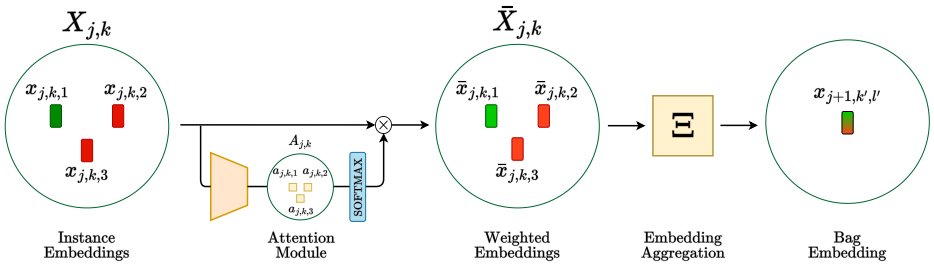


Figure 5.2: MIA block. Input embeddings are fed into the attention module to compute attention scores. Green instances represent a positive class, while a red represents a negative class. Attention scores are used to compute weighted representations of the instance embeddings. The weighted embeddings are thereafter aggregated to create a final bag embedding.

and computes attention scores that leverage the meaningfulness of the features extracted for the given task by learnable parameters. Finally, the attention scores are aggregated to obtain the bag representation.

For ease of exposition, we define indexes $\delta = (j, k, l)$, $\delta' = (j + 1, k', l')$, $\gamma = (j, k)$. We omit the use of training sample index i. An attention score a_δ for a given input embedding x_δ is calculated as:

$$a_\delta = \exp\{\mathbf{w}^\top(\tanh(\mathbf{V}\mathbf{x}_\delta^\top) \odot \text{sigm}(\mathbf{U}\mathbf{x}_\delta^\top))\} \tag{5.6}$$

where $\mathbf{w} \in \mathbb{R}^{L\times 1}$, $\mathbf{V} \in \mathbb{R}^{L\times M}$ and $\mathbf{U} \in \mathbb{R}^{L\times M}$ are trainable parameters and $\odot$ is an element-wise multiplication. Furthermore, the hyperbolic tangent $\tanh(\cdot)$ and sigmoid $\text{sigm}(\cdot)$ are included to introduce non-linearity for learning complex applications. The attention scores a_δ are normalized into $\tilde{a}_\delta$ to ensure that the sum of the components of the attention scores vector is 1, as this makes it possible to have variable bag sizes.

$$\tilde{a}_\delta = \frac{a_\delta}{\sum_{l=1}^{L_\gamma} a_\delta} \tag{5.7}$$

Note that the unnormalized a_δ would better reflect the attention score directly, and we use that for visualization in the experiments. Finally, the aggregation function $\Xi(\cdot)$ transforms a leveraged bag of embeddings $\tilde{\mathbf{X}}_\gamma$ to obtain a bag representation $\mathbf{x}_{\delta'}$ as:

$$\mathbf{x}_{\delta'} = \Xi(\tilde{\mathbf{X}}_\gamma) = \Xi(\{\tilde{a}_\delta \cdot \mathbf{x}_\delta, \forall l = 1, ..., L_\gamma\}) \tag{5.8}$$

The proposed neural network architecture NMIA combines a recursive bag processing of low-dimensional embeddings and attention mechanisms. The multiple instance with attention (MIA) block computes attention scores for each input embedding and aggregates them to create an embedding that represents the bag's contents, see Figure 5.2. Instances are transformed into embeddings, weighted and aggregated into a bag-of-bags embedding after passing through the MIA blocks. As a result, a bag-of-bags embedding $\mathbf{x}_{J+1,1,1}$ is formed. Finally, a classifier Θ_c predicts the label $\hat{y}$ for the sample $\mathbf{X}$. The model's forward propagation is summarized in Algorithm 2. Although the example given corresponds to features extracted from images, this problem can be extrapolated to other signal types that can be represented as feature embeddings.

Algorithm 2 NMIA Forward Propagation

Input: image bag-of-bags sample $\bar{\mathbf{X}} = \{\bar{\mathbf{x}}_l, \forall l = 1, ..., L\}$
Output: classifier prediction $\hat{y}$
  / Transform input image instances into feature embeddings*/*
  for $l = 1$ **to** L **do**
    $\bar{\mathbf{x}}_l = G_f(\mathbf{x}_l)$
  end for
  for $j = 1$ **to** J **do**
    for $k = 1$ **to** K_j **do**
      for $l = 1$ **to** $L_{k,j}$ **do**
        */*Define indexes* $\delta = (j, k, l)$, $\delta' = (j + 1, k', l')$**/*
        $a_\delta = \exp\{\mathbf{w}^\top(\tanh(\mathbf{V}\mathbf{x}_\delta^\top) \odot \text{sigm}(\mathbf{U}\mathbf{x}_\delta^\top))\}$
      end for
      $\tilde{a}_\delta = \frac{a_\delta}{\sum_{l=1}^{L_{k,j}} a_\delta}$
      $\mathbf{x}_{\delta'} = \sum_{l=1}^{L_{k,j}} \tilde{\mathbf{x}}_\delta = \sum_{l=1}^{L_{k,j}} \mathbf{x}_\delta \tilde{a}_\delta$
    end for
  end for
  */*Resulting bag-of-bags embedding is fed into the classifier*/*
  $\hat{y} = \Theta_c(\mathbf{x}_{J+1,1,1})$
  return $\hat{y}$

We conduct an experiment where at least two instances from the same inner-bag have to be positive for the weak bag-of-bags label to be positive, as described in Eq. (5.9). we have considered the positive instance class y_{l+}^1 as the digit 9 for MNIST and tiles with metastatic tissue for PCAM. This

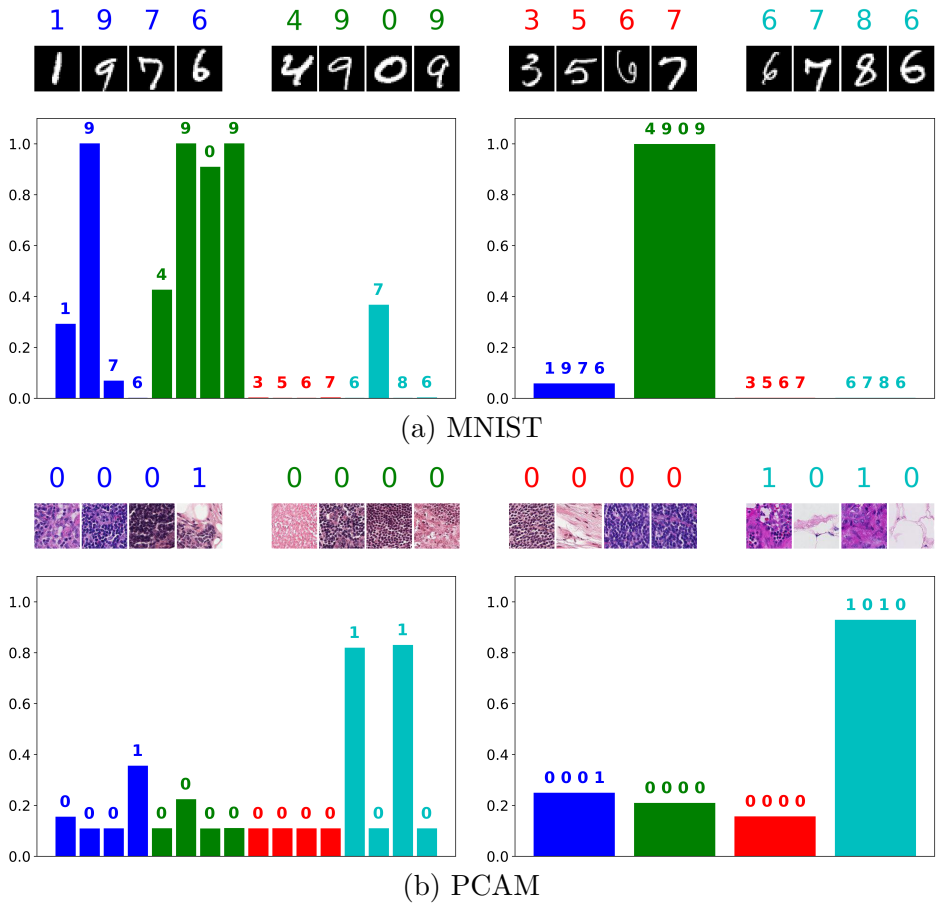
is motivated from region-based analysis of medical images, where typically an object belonging to a positive class is located in a specific region and not scattered across the entire image. Therefore, this particular positive region of the image will contain several positive instances. Regions containing few positives are regarded as noise or misclassified instances, and they should not be reflected in the overall prediction.

$$y_l^2 = \begin{cases} 0, & \#y_{l^+}^1 \leq 1 \\ 1, & \#y_{l^+}^1 > 1 \end{cases} \quad \overset{\mathbf{1^{st}}}{} \qquad y = \begin{cases} 0, & y_l^2 \notin 1 \\ 1, & y_l^2 \in 1 \end{cases} \quad \overset{\mathbf{2^{nd}}}{} \tag{5.9}$$

Implementing a model with an attention mechanism provides an edge over the model that does not in terms of interpretability, as shown in Figure 5.3. We can observe that the attention mechanism can correctly identify the positive instances at the instance level and recognise the positive inner-bags, distinguishing them from those containing noisy instances. A conventional weakly supervised model would not perceive this sense of location since all instances are encapsulated under the same bag. This results in more interpretable predictions on account of which instances and inner-bags the model weighs more.

Summary of Contributions

This paper proposes the NMIA architecture with the aim of overcoming intricate data dependencies for weakly supervised scenarios. A generalized notation for the implementation of nesting for any number of levels is formulated. NMIA grants higher flexibility and performance in comparison to non-nested architectures.



(a) MNIST

(b) PCAM

Figure 5.3: Examples of attention scores a_δ in test samples with 2-level NMIA for (a) MNIST and (b) PCAM. A single positive instance in the inner-bag is considered noise whereas two or more should give a positive inner-bag, resulting in a positive bag-of-bags bag. Instances are categorized by colors indicating inner-bag belonging, while digits represent the true label of the instance. Bar plots on the left column show the attention at level 1, while on the right, at level 2. Positive instances obtain the highest attention scores.

Chapter 6

Region of Interest Extraction

The following chapter covers the complex process of extracting regions of interest (ROIs) from WSIs. We explore the methodologies, techniques, and challenges associated with this crucial step. We aim to unravel the significance of precise ROI extraction. We compare the impact of manually annotated and automatically segmented areas on deep learning models performance. We also investigate data-efficient approaches to minimize manual labor. Paper 2 is dedicated to the topic, and the main methods, results, and contributions will be presented. In the thesis, Paper 2 is part of sub-objective SO_1, on utilizing non-supervised learning approaches to address the lack of annotations, and SO_2, for enabling the implementation of automated deep learning pipelines.

6.1 Paper 2 - Tissue Segmentation

In this section, the contributions in Paper 2 are presented. This paper presents an active learning-based method for domain adaptation of a WSI tissue segmentation model, as highlighted in Fig. 6.1. The proposed active learning method is explained in 5.1. The dataset utilized corresponds to Dataset A, as listed in Chapter 4.3. Tiles were extracted at 25x, 100x and 400x magnifications, thus forming a triplet, and a maximum of 500 triplets were extracted per WSI. A tile size of 128×128 was used, and all three tiles in a triplet share a common physical point as the center pixel.

We compare the performance of the model pre-adaptation TRI-CNN, where we obtain the trained model [26], to models trained with our proposed

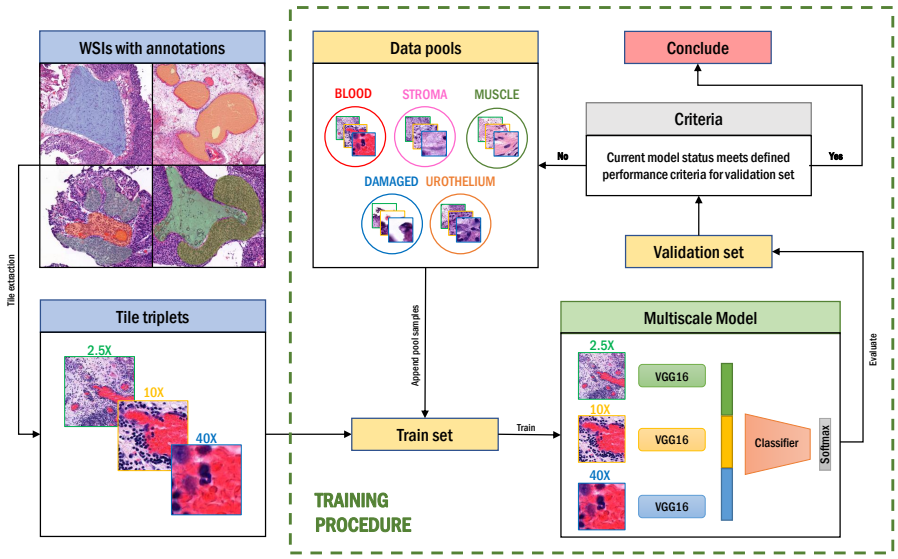


Figure 6.1: Active learning framework for training a model TRI-AL. Pathologists annotate ROIs, from which tile triples are extracted at different magnifications levels (25x, 100x, 400x) to form a triplet. During the training stage, a starting training set is defined for training a multiscale model. The model performance is evaluated on the validation set. Then, stopping criteria decides to resume with another training iteration or conclude the learning. In case that the criteria is not met, new samples from the pools of data are drawn and appended to the current version of the train set.

active learning strategy to train a model TRI-AL. The initial training set $\mathcal{X}_0$ is made of 25000 triplets per class. At every iteration i, 20000 new samples S from the pool $\mathcal{P}$ are added to the training data set $\mathcal{X}_j$. We defined a maximum number of iterations for training the model TRI-AL$_{\text{ITER}}$.

The development of the training set class distribution for the active learning-trained model TRI-AL$_{\text{ITER}}$ in comparison to fixed distribution of the training set $\mathcal{X}$ is shown in Fig. 6.2. Looking at the relative data balance per class, we identify a tendency over the iterations where damaged, muscle and urothelium class distribution δ reach an asymptote. Also, we observe that the algorithm deems to suggest that stroma is significantly harder to learn and requires a more extensive number of data points in comparison to other classes. Even then, the margin for δ progressively slows down and we would expect it to become stagnant. By evaluating the progression of class balance, we can propose more effective annotation guidelines. This approach would accurately indicate pathologists the specific annotations

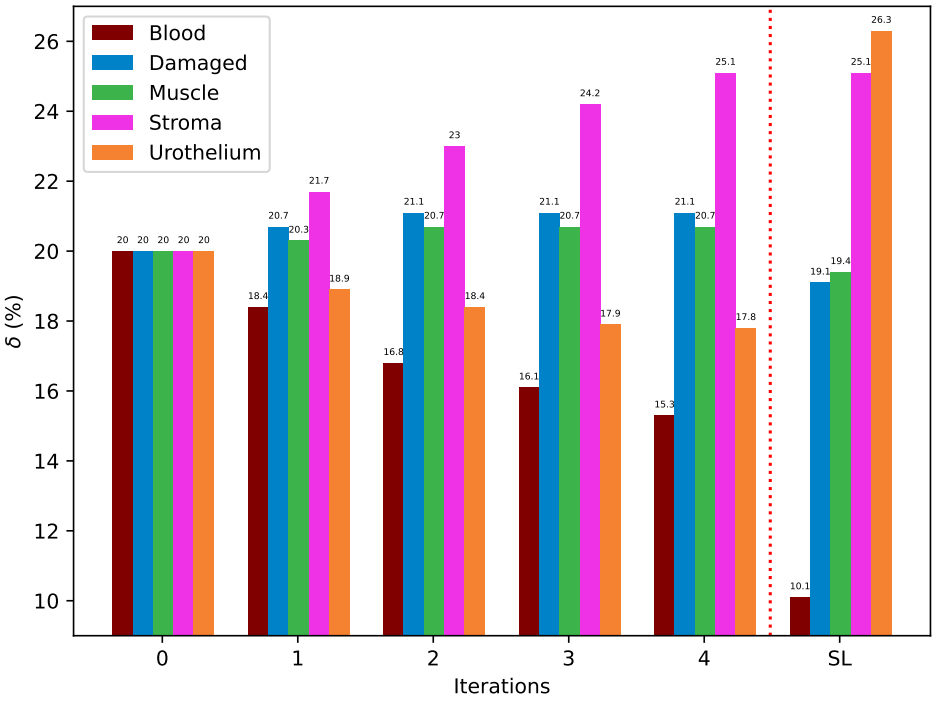


Figure 6.2: Class relative size per class δ_i over the iterations for TRI-AL$_{\text{ITER}}$ model. We observe that the algorithm prioritizes harder classes over simpler ones as the iterations pass. As $\mathcal{X}_j$ increases, δ_i reaches an asymptote for damaged, muscle and urothelium tissue classes, while penalizing blood in favor of stroma. Class distribution does not match that annotated from a pathologist, as per the SL section.

required for deep learning models, allowing them to efficiently invest time before committing to substantial efforts.

Pathologists also visually inspected the model's segmentation results by overlaying the predicted masks over the raw WSIs. Illustrative examples of the model's segmentation results for a representative WSI of the test set are shown in Fig. 12.3 for a ROI and the entire WSI, respectively. Upon visual inspection of the results by experts, according to TRI-CNN segmentation, we have identified four main observations: staining effects leading to false positives of blood in regions with high levels of eosin stain, non-cauterized damaged areas such as blur or folding, the risk of misinterpreting infiltrative immune cells as urothelial cells, and the potential for the model to predict urothelium with significant cytoplasm as stroma. As per TRI-AL$_{\text{ITER}}$, it was confirmed that the model accurately segmented and classified different

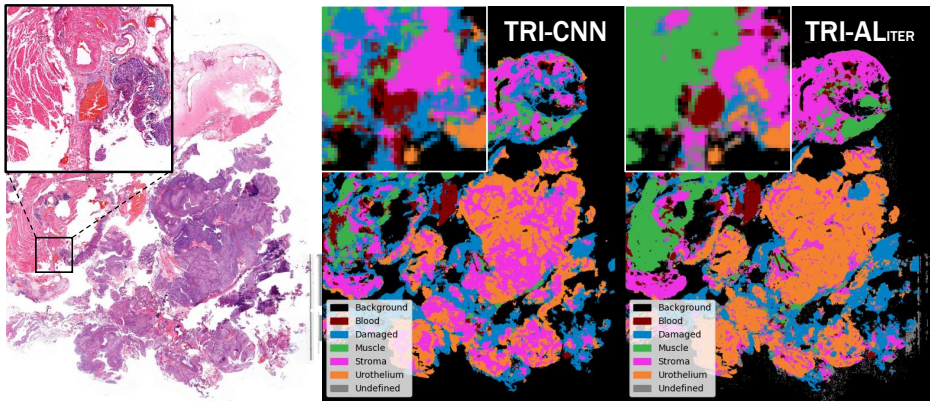


Figure 6.3: Segmentation masks of a WSI. Areas from all tissue types are extracted for analysis at the inference stage, resulting in a labelled colormap based on tile predictions. The active learning model TRI-AL$_{\text{ITER}}$ yields a more accurate segmentation than the pre-trained model TRI-CNN.

tissue types. In comparison to supervised learning (SL) strategies, active learning strategies excel in achieving higher performance by utilizing data more efficiently, thereby demonstrating their ability to achieve comparable or superior results while requiring less data for training.

Summary of Contributions

This paper proposes an active learning framework with a multi-scale CNN for domain adaptation of a tissue segmentation model of bladder cancer histopathological images. This method outperforms supervised learning strategies using a limited fraction of the training set. The method also serves as suggested annotation guidelines based on class distribution demand.

6.2 Exploiting Domain Knowledge using Segmentation Maps

Exploiting domain knowledge through segmentation maps is a pivotal aspect of the CPATH framework in computer-aided pathology and the work presented in this thesis. These maps provide detailed annotations of WSIs, delineating ROIs and aiding in the interpretation of medical data.

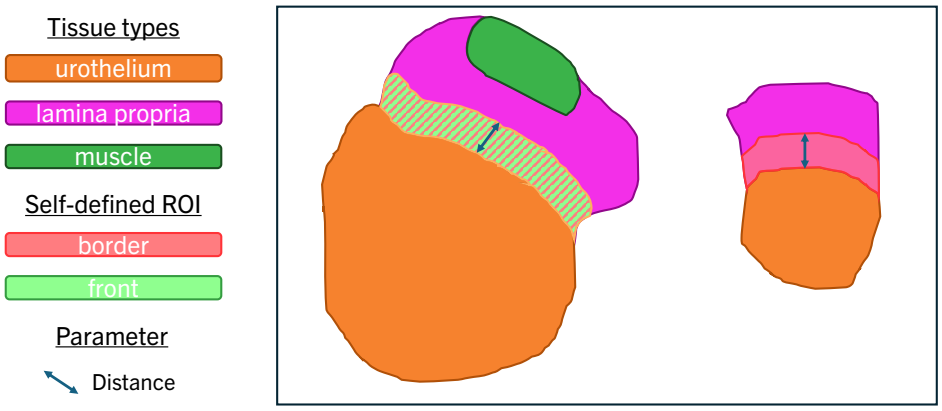


Figure 6.4: Schematic representation of ROI definition based on tissue segmentation maps. Given a segmentation mask with the listed tissue types, we define ROIs automatically using morphological operations. In the example presented, we extract the border area between urothelium and lamina propria applying dilation and a distance limitation determined by domain knowledge from expert pathologists. We further delineate a subset of the borders by extracting only those areas where muscle is present within the same tissue section, aiming to represent the potential invasive front.

One of the primary benefits of segmentation maps within CPATH is their role in enhancing the performance and interpretability of other model predictions. By visually representing pathological features such as tissue types, these maps provide precise annotations for enabling practitioners to contextualize model outputs within the intricate anatomical structures present in WSIs. Moreover, segmentation maps play a crucial role in guiding feature extraction within the CPATH framework. By highlighting relevant regions, they assist in the identification of discriminative features associated with specific pathologies, thereby empowering other models to make more accurate predictions. Furthermore, segmentation maps can support transfer learning by serving as auxiliary tasks for pretraining on related datasets. This strategic approach allows the model to leverage knowledge from external sources, enhancing its adaptability across diverse pathology datasets and expanding its potential for real-world applications.

Building upon the previously mentioned advantages, this thesis capitalizes on the segmentation maps obtained to enhance tile extraction strategies. This enables the exclusion of irrelevant tissue areas, which may introduce noise to the application under development, allowing us to concentrate solely on the designated ROI. To define and extract tiles from ROIs, we

employ automated image processing techniques such as morphological operations, following insights derived from domain experts. This concept is illustrated in Figure 6.4. Segmentation mapping produces tissue masks outputs, including lamina propria, urothelium, and muscle. However, determining tumor proximity areas, such as the bordering regions between lamina propria and cancerous urothelium, or selecting regions adjacent to muscle tissue, relies on identifying the invasive front, which in turn depends on domain-specific knowledge. To automatically obtain these self-defined regions, as delineated in the two aforementioned examples, we utilize a disk dilation morphological operation on the urothelium and lamina propria masks, determined by the desired physical depth and pixel size. Then, overlapping areas are segmented for tile extraction within the newly delineated bordering region. Alternatively, we can refine this border ROI by employing a region-growing algorithm along these borders to eliminate areas lacking muscle tissue within a tissue section.

Chapter 7

Diagnostics

In this chapter, we present diagnostic applications for NMIBC WSIs, specifically focusing on the grading and staging. We aim to highlight the significance of these applications in advancing precision diagnostics for improved patient outcomes. Papers 3 and 4 are dedicated to the topic, and the main methods, results, and contributions will be presented. In the thesis, Papers 3 and 4 are part of sub-objective SO_2 on the development of automated deep learning methods for classifying risk factors based on histological features.

7.1 Contributions overview

The primary objective is to underscore the substantial contributions of deep learning methodologies in advancing precision diagnostics for NMIBC WSIs. Specifically, the implementation of fully automated deep learning pipelines for grading and staging (see Section 3.2). By doing so, we aim to offer more efficient tools for the assessment of NMIBC, optimizing the overall pathology practice while enhancing patient management decisions. Accurate grading and staging of the tumor is important to categorize patients into risk groups, and guides which treatment strategy the patient will receive, and, thus, patient outcome.

Automated systems capable of identifying critical malignant areas can aid uropathologists by diminishing the effort required to provide accurate diagnostic reports. Given the high heterogeneity of bladder cancer, it is relevant to interpret visual local patterns based on contextual neighbouring

regions, thus a pathologist would zoom in and out to get an understanding of both cellular and morphological details. Similarly, the performance of a DL model is expected to improve when patches from various magnification levels are provided as input to the system. Furthermore, it is important to leverage the impact of manual annotations for supervised models and self-supervised learning methods, which operate without labels or rely solely on overall patient labels. Although annotations grant ground truth, they are time-consuming and limited to expert input efforts, rendering supervised methods dependent and costly. However, by using automated methods that employ non-supervised methods and clinical labels exclusively, eliminates the requirement of manual input and enables applications to autonomously analyze raw data and produce results during the inference stage.

7.2 Paper 3 - Invasive Cancerous Areas

In this section, the contributions in Paper 3 are presented. This paper presents a CNN-based deep learning algorithm that can process all available tissue areas in a WSI and produces a heatmap to detect invasive cancerous areas (ICA). The algorithm finds ICA associated with T1 stage across the WSI without the need of previous ROI or tumour segmentation. The dataset utilized corresponds to Dataset B, as listed in Chapter 4.3. Multiple models using different combinations of magnification levels were trained and tested. The following combinations of magnification views were used: MONO (400x), DI (400x, 100x), and TRI-scale (400x, 100x, 25x). Tiles of size 256×256 were extracted from annotated regions at 400x magnification, utilized for training purposes. Tiles at different magnifications were extracted so that the center of the tile remains the same for all views. Extracted tiles at 400x are required to be covered by at least 70% of a tissue mask in order to be assigned such label. From the extracted tiles, we defined two subsets; i) a *control set*, and ii) an *inclusion set*. On the one hand, a control set is sampled, containing ICA and selected non-invasive areas (*Uro*, *LP*). Extracting carefully selected tiles, we remove the surrounding tissue present in the slides while maintaining tiles related to the ROI. On the other hand, the inclusion set encompasses all tissue present in a WSI, adding non-diagnostically relevant tissue. The objective of considering all tissue simultaneously is to manage ROI definition and invasion detection automatically. This ensures that during the inference stage, we are not reliant on a separate algorithm for ROI definition.

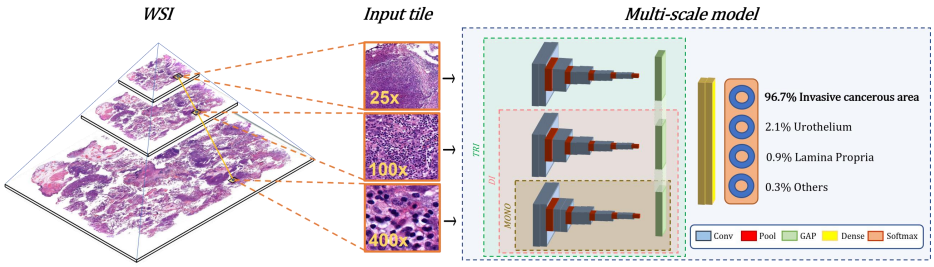


Figure 7.1: Input tiles at different magnification levels are extracted from the WSI and fed into the feature extractors. Extracted features are concatenated and fed into the dense layers which will give the final prediction.

The algorithm consists of a combination of VGG16 backbones with input tiles at different magnification levels to obtain detailed as well as contextual information. This information is later concatenated to embeddings from other magnification views and fed into the dense layers which will give the final prediction, as shown in Fig. 7.1.

For the first set of experiments, denoted E_{mag}, MONO, DI and TRI-scale models are evaluated to determine the relevance of contextual tissue morphology for discerning invasion. TRI-scale models provide the best results for ICA detection, inferring that regional context derived from tissue morphology is beneficial in the task of discerning invasive patterns over non-invasive ones. Hence, models that incorporate lower magnification views provide better performance than those who focus on local patterns solely.

For the second set of experiments, denoted E_{ds}, we ascertain the performance variation for adding non-ROI tissue. Results on the control set demonstrate that discerning invasive from non-invasive patterns is possible. We observe that models trained on the inclusion set provide comparable, but lower, results. It is, however, with models trained on the inclusion set that we can directly deploy the algorithm in an inference pipeline for WSI analysis. The best performing model of the inclusion set, TRI_{incl}, was used to produce probability heatmaps over WSIs from the test set, as shown in Fig. 7.2. To simplify the visualization, ICA was designated as the positive class, while all other tissues were grouped into a single negative class.

For the third set of experiments, denoted E_{tils}, we excluded all tiles which included *"tumour infiltrating lymphocytes"* (TILs). TILs are small immune cells that can be found either in the urothelium or the lamina

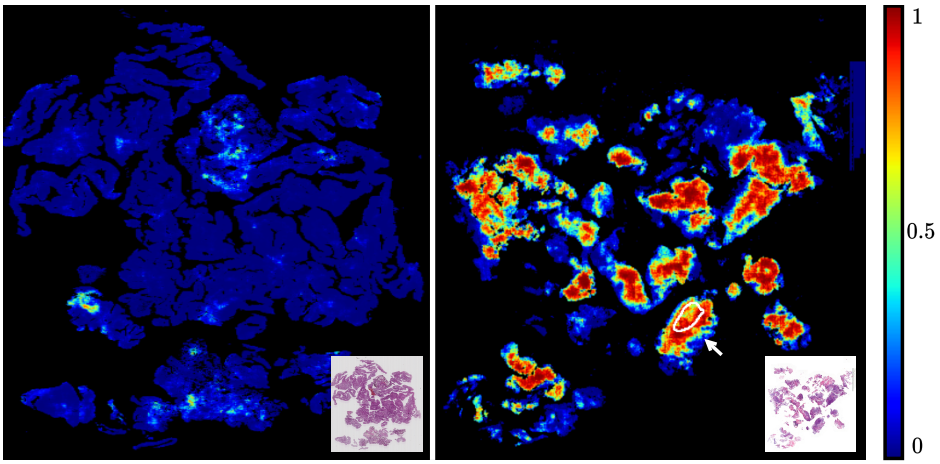


Figure 7.2: Heatmaps showing the probability of a patch to belong to the ICA class, based on the TRI_{incl} model. Two test WSI samples with stage Ta (left) and T1 (right) are shown. An annotated ICA region is highlighted by a white border in the T1 WSI. Inclusion set models support the analysis of the entire WSI, not limited to pre-defined ROIs. A thumbnail of the original WSI is displayed at the bottom right corner.

propria. TILs may lead to confusion for discerning them from invasion. We noticed that the presence of TILs dampers the predictive performance, as models struggle to differentiate such areas from ICA. Tiles where TILs are present are often missclassified for ICA.

Finally, for the fourth set of experiments, denoted E_{emc}, we used tiles from one of the hospitals exclusively. We chose EMC over SUH since the number of available data from EMC is far greater, especially regarding the number of ICA tiles. E_{emc} will show how impactful data balance among classes and domain shift are toward detecting invasive cancerous patterns. Also, the experiments were run in a binary and multi-class manner to assess whether grouping of non-invasive tissue improves ICA detection performance. Despite reducing scanning and staining variability, performance was poorer in comparison to the other experiments. This can be due to the lack of a substantial number of extracted urothelium tiles, which are present mostly in the excluded SUH cohort.

Summary of Contributions

This paper proposes multi-scale CNN based models for detecting ICA across the WSI. Eliminating non-diagnostically relevant regions barely increases

model performance. Automatic ICA detection sets the first step towards automatic staging of NMIBC tumors, as the presence of ICA is indicative of higher stages.

7.3 Paper 4 - Pathological Grade

In this section, the contributions in Paper 4 are presented. This paper presents a new pipeline for grading NMIBC WSIs named NMGrad. The pipeline involves tissue segmentation, categorization of urothelium areas into location-dependent regions, and employs weakly supervised learning and attention mechanisms for grading, resulting in innovative diagnostic suggestions with heatmaps highlighting tiles and ROIs. The dataset utilized correspond to Dataset C, as listed in Chapter 4.3. As we employ a multi-scale CNN input, various magnification are used. In the case of mono-scale models MONO, we employed 400x magnification. For di-scale models DI, we used 400x and 100x. Finally, for tri-scale models TRI, we used all three available magnification levels (400x, 100x, and 25x, organizing them into triplets by bundling the corresponding tiles. We do, however, maintain a consistent tile size of 128×128 across magnifications.

We introduce a three-step fully-automated pipeline that combines lesion segmentation, ROI definition and multiple instance learning (MIL) for predicting the WSI-level grade, as depicted in Fig 7.3. First, the urothelium lesion of a raw WSI is segmented using a pre-trained algorithm. Next, employing the segmented urothelium mask, we divide it into regions by implementing morphological operations to eliminate smaller areas. Subsequently, we cluster larger regions based on image coordinates, ensuring comparable region sizes. Then, tiles are extracted from the defined regions to train MIL models for WSI grading.

We explored the significance of magnification-embedded information, by comparing MONO, DI and TRI-scale models. We promptly discovered that increasing the number of input magnifications enhances overall performance, resulting in TRI-scale models to yield the best results. We also compared various aggregation techniques in weakly supervised learning, including the NMIA architecture proposed in Paper 1. Utilizing NMIA resulted in a notable performance improvement, particularly in enhancing performance when analyzing scattered tissue regions across the WSIs. An example of the region attention score heatmap is depicted in Fig. 7.4. The analysis reveals that low-grade instances tend to have lower attention scores and

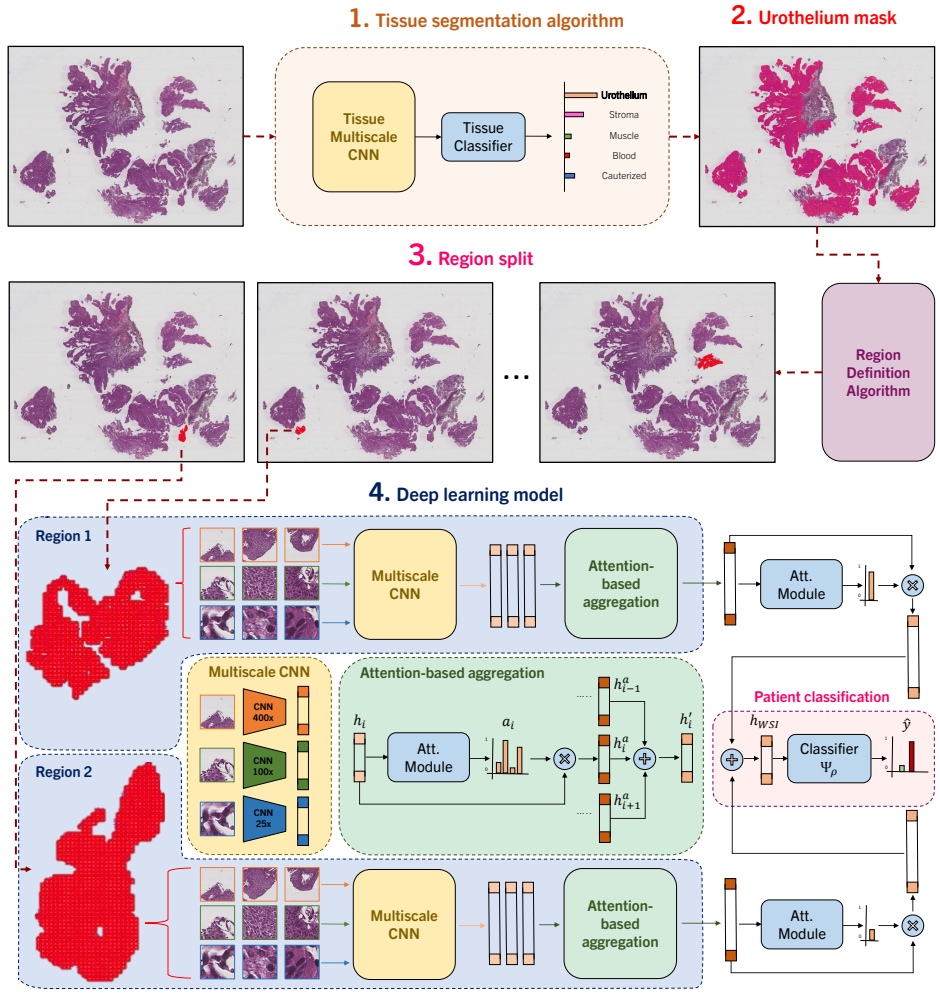


Figure 7.3: NMGrad pipeline. Initially, we apply a tissue segmentation algorithm for ROI extraction Then, we pinpoint diagnostically significant urothelium areas within WSIs. Subsequently, we split the urothelium mask into regions, based on proximity and size, and extract tile triplets. In a hierarchical fashion, we further transform these triplets within their corresponding regions into region feature embeddings using an attention-based aggregation method. All the region representations are then consolidated into a comprehensive WSI-level representation through a weight-independent attention module. Finally, this WSI feature embedding is input into the WHO04 grading classifier in order to produce accurate WSI grade predictions.

prediction outputs, while high-grade instances exhibit the opposite. This

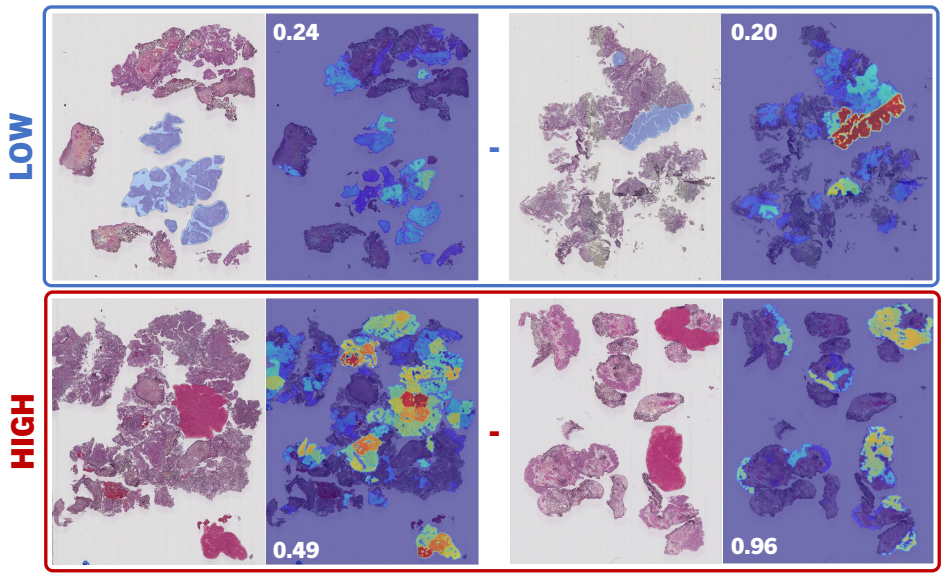


Figure 7.4: Region-level attention score heatmaps. Example regions of annotations of low- and high-grade ROIs annotated by an uropathologist are compared to the output attention provided by the proposed model NMGrad, left to right respectively. The choice of annotated ROIs correspond to highest attention scores, for red and blue correspond to low and high attention correspondingly. We have included the WSI-level prediction score for reference.

aligns with observations indicating that positive high-grade instances have more focused attention scores, as the positive class is more meaningful than the negative for MIL. Interestingly, misclassified WSIs show distinct characteristics, with low-grade slides displaying both high attention and prediction scores, and high-grade slides exhibiting a broader range of values. Additionally, regression lines for true positives and false positives, as well as true negatives and false negatives, display similar trends, suggesting consistent misprediction characteristics across different classes.

We enhance the evaluation process by incorporating correlation calculations with follow-up information, ensuring a comprehensive assessment of our model's performance. These correlations indicate that NMGrad's grade may offer greater predictive value in assessing the likelihood of progression in NMIBC, with no correlation found for recurrence as anticipated.

Summary of Contributions

This paper proposes a fully-automated framework for grading NMIBC WSIs. We integrate a tissue segmentation algorithm to outline the urothelium. Additionally, we utilize morphological and clustering operations to define ROIs. Ultimately, leveraging multi-magnification inputs and nested model architectures, we deliver a grade prediction at the WSI-level. Clinical evaluations suggest that our model outperforms previous state-of-the-art methods, while yielding statistically significant correlations with clinical outcomes.

Chapter 8

Prognostics

This chapter is dedicated to prognostic prediction of clinical outcome of NMIBC patients using histological WSIs. Paper 5 is dedicated to the topic, and the main methods, results, and contributions will be presented. In the thesis, Paper 5 answers the sub-objective SO_3 on the development of clinical outcome prediction from NMIBC WSIs.

8.1 Paper 5 - Prognostic Prediction

In this section, the contributions in Paper 5 are presented. This paper presents a pioneering investigation into the application of deep learning techniques to analyze histopathological images for addressing the substantial challenge of automated prognostic prediction. The prognostic applications covered are recurrence and treatment outcome prediction. The treatment corresponds to BCG, and the usage of deep learning for BCG response prediction using WSI has not been previously explored. The datasets utilized correspond to Dataset C and D, as listed in Chapter 4.3. Various magnification levels and tile sizes are used. In the case of mono-scale models, we used a tile size of 256×256 for 100x magnification and 512×512 for 200x magnification. This decision was made to ensure that the patches covered the same physical area, i.e. field of view. In the case of multi-scale models TRI, we used a tile size of 128×128 for all three magnification levels (400x, 100x, and 25x).

We introduce a three-step fully-automated pipeline for WSI prognosis that combines region of interest (ROI) extraction, contrastive learning

for feature representation and multiple instance learning (MIL) for predicting the concluding prognostic outcome, as depicted in Fig 8.1. This approach enables us to optimize the model by leveraging the benefits of these techniques. It ensures the inclusion of important instances for predicting clinical outcome from WSI visual cues, while maintaining computational feasibility. Ultimately, the proposed steps for prognostic predictions in histopathological imaging are the following:

A Define and extract ROIs based on clinical domain knowledge using a tissue segmentation algorithm for tile extraction strategies.

B Train a feature extractor G_θ to generate an intermediate dataset $\mathcal{H}$ using contrastive learning.

C Use image feature embeddings $\mathcal{H}$ for prognostic classification using MIL.

We explored the significance of various ROI configurations, guided by clinical domain knowledge, as the localization of the tissue of interest is unknown. We employed tissue segmentation algorithms, from Paper 2, for extracting tissue masks and define a ROI D. Ultimately, results indicate that the union of areas of urothelium and lamina propria are the most informative. We also explored the relevance of multi-scale models, however, utilizing a fixed magnification consistently yields highest performance.

We investigated the importance of various feature extraction strategies, by comparing CNN architectures and contrastive learning methods. We found that DenseNet121, with ImageNet pre-trained weights, coupled with self-supervised contrastive learning produces the best results.

We conducted a comparison between different aggregation techniques in weakly supervised learning, including the architecture proposed in Paper 1, NMIA. The utilization of attention-based models leads to a notable performance improvement. Moreover, when examining the scattered tissue regions across the WSI using NMIA, we observe the most significant performance enhancement across all techniques. An example of attention score heatmap is visualized in Fig. 8.2. Pathologists often use tissue punching to select clinically relevant regions for subsequent analysis in tissue microarrays. The model tends to allocate its highest attention to punched areas, and it is crucial to acknowledge that diagnostically relevant information may extend beyond the punched areas. The alignment between the model's attention

focus and the clinical practice of region selection through punching is an encouraging and promising finding.

At the inference stage, a fully-independent system cannot be conditioned by manually annotated regions. Even at the training stage, we have observed the challenges in obtaining annotations for all WSI due to the labor-intensive nature of the process. For a prognostic task, the relevance of manual annotations remains uncertain. To explore this aspect, we conducted an experiment where the model was trained on annotated regions, but tested on automatically segmented regions, and vice versa. This was compared to a fully automated system for both training and inference. Generally, we observed that the models perform better after using auto-generated ROIs in the training and annotations for testing, which we interpret as the models benefiting from the larger dataset that is available when we can include non-annotated data.

Summary of Contributions

This paper proposes a three-step automated pipeline for the challenging task of prognostic prediction, eliminating the need for manual annotations. We explore the relation between ROIs and patient outcomes. The utilization of deep learning for BCG response prediction using WSI is pioneer, although achieving fully automated prognostics based solely on WSI remains a challenging task.

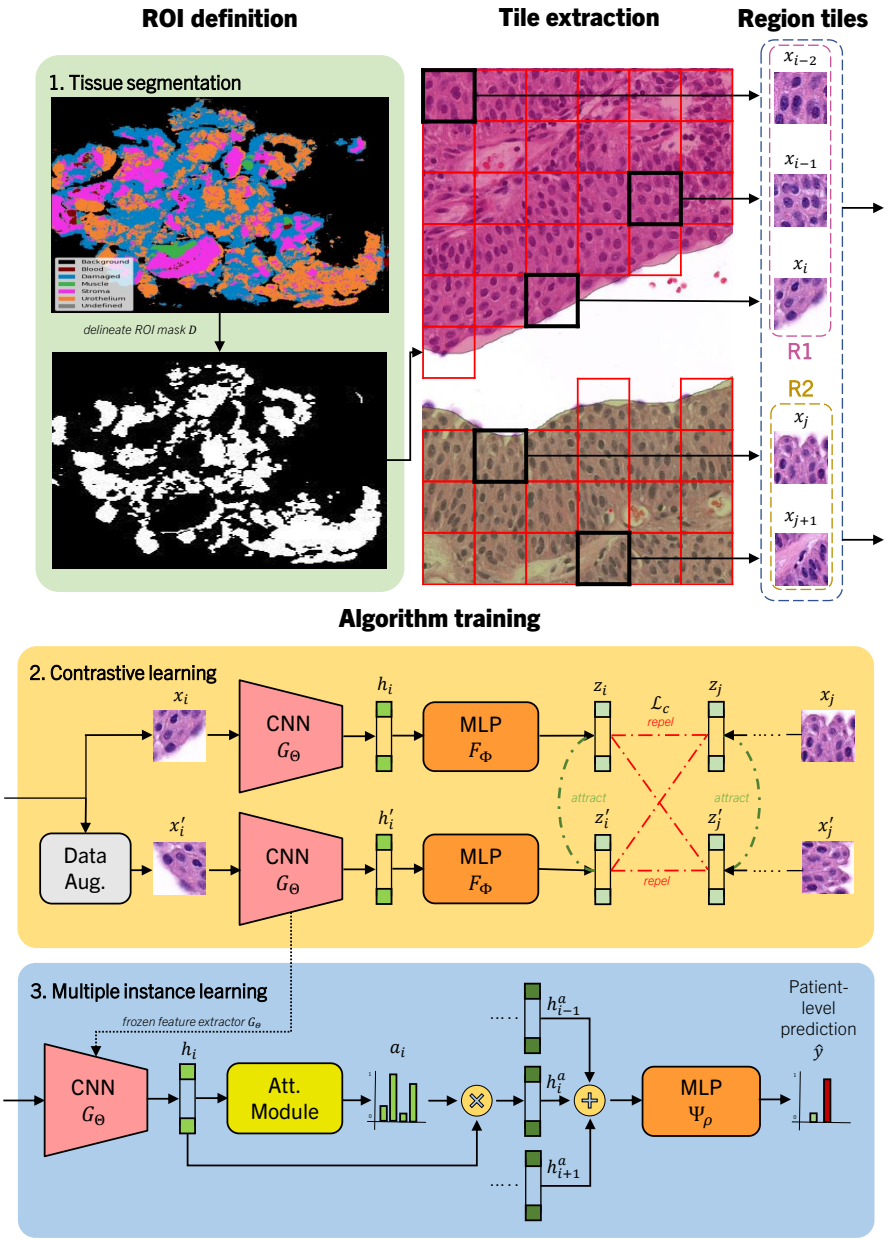


Figure 8.1: Deep learning pipeline for prognostic outcome prediction. 1) A tissue segmentation is employed for delineating a ROI of choice D. Then,tiles are extracted from WSI regions for training an algorithm. 2) Contrastive learning is employed to learn representations of the tiles. 3) The representations are then used to train an AbMIL model that predicts the prognostic outcome.

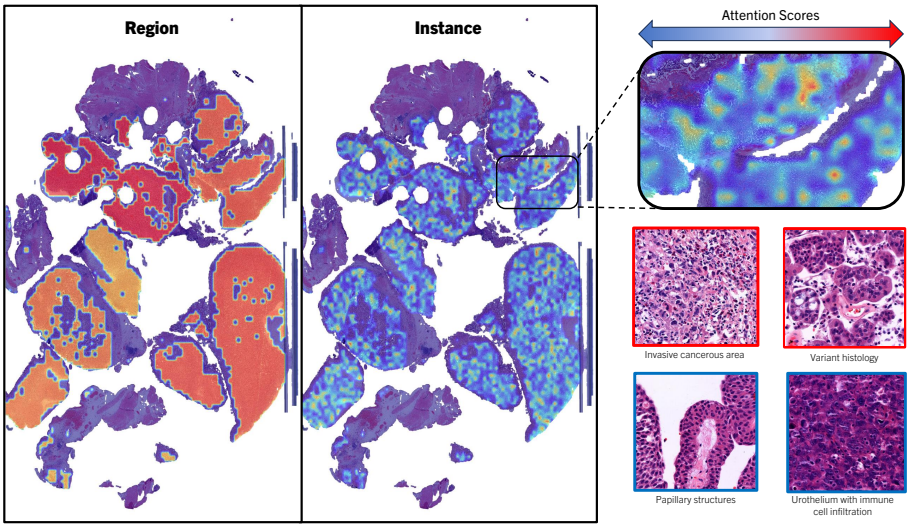


Figure 8.2: Heatmap illustrating attention scores over a WSI. The heatmap provides insights into the ROIs where the attention is concentrated within the WSI, facilitating a better understanding of prediction dynamics and highlighting areas of significance for clinical interpretation.

Chapter 9

Discussion and Conclusion

This chapter summarizes how the contributions of this thesis lead to the achievement of the established objectives, discusses the encountered limitations, and, consequently outlines avenues for future work.

9.1 Discussion

In the contemporary landscape of ML the demand for adaptable methodologies that transcend disciplinary boundaries is paramount. This study explores the development of versatile methods, poised to extend beyond CPATH and find application across diverse domains. By addressing broader challenges in learning from data, we aim to drive innovation in resource-efficient learning techniques.

In the context of NMIBC application, the stratification of patients into risk groups according to EAU guidelines is crucial for predicting clinical outcomes [13, 96]. Quantifying risk factors is highly relevant as it allows for objective assessment and comparison of patient data. This quantitative approach provides clinicians with precise information to stratify patients into appropriate risk groups, aiding in treatment planning and prognostication [8, 14–17, 45, 123]. Therefore, algorithms capable of providing quantitative results for these factors are essential. Additionally, since annotating WSIs is a laborious and time-consuming task, we explore methodologies aimed at reducing reliance on extensively annotated data and expert pathologist opinions [46–49]. This allows for more efficient and scalable analysis of WSIs, ultimately improving patient risk stratification and clinical decision-making. Thus, the thesis explored the development of fully automated

pipelines for analyzing NMIBC WSI to offer quantitative and clinically relevant disease assessment.

9.1.1 Label-efficient Algorithms

This section is directly related to the thesis sub-objective SO_1: "Explore the utilization of non-supervised learning approaches to address the absence of annotations".

Given the limited availability of annotated data in the medical field, meeting the substantial requirements for annotated regions in supervised learning networks can be challenging. In light of the promising avenues observed with various approaches [84, 134, 166–174], we investigate alternative learning strategies. Moreover, we inquire in the development of tailored application algorithms.

In Paper 1, a novel ML architecture was presented, referred to as NMIA. The proposed algorithm incorporates the capability to adapt to nuanced data structures for categorized data processing. The model is trained using solely a weak label in a MIL fashion. Coupling attention mechanisms proved to be beneficial, both for enhancing interpretability and information integration, as well as for improving model performance. NMIA was also adopted in Paper 4 and Paper 5, wherein the architecture exhibits superior predictive capabilities in comparison with related weakly supervised architectures.

Paper 2 presents an active learning framework for efficient domain adaptation of DL algorithms in the CPATH domain. The prototypical task presented is that of tissue segmentation of NMIBC WSIs. The framework provides an intuitive description of the model learning process, and consequently iteratively adapts based on the current predictive faults on the validation set. Tracking the evolution of the model needs permits the user to tailor economic annotation strategies. Results have shown that the proposed method is more label-efficient than a supervised learning approach, in conjunction with overall major accuracy.

9.1.2 Automated Histopathological Analysis

The following section is connected to the thesis sub-objective SO_2: "Create automated deep learning systems for classifying risk factors based on histological features".

Due to the vast size of WSIs and the presence of undesired tissue and artifacts, histopathological image analysis has historically relied on carefully selected ROIs, often requiring manual input from pathologists to delimit the area of interest [34, 38]. Efforts toward fully-automated pipelines for WSIs analysis have been made, emphasizing on lesion segmentation as a preliminary step [31, 33, 161, 175]. However, clinical applications on NMIBC remain largely unexplored, particularly concerning pathologist-independent algorithms.

Regarding the evaluation of clinical risk factors for NMIBC, we identified challenges in the visual interpretation of histopathological features in WSIs [176, 177]. The analysis of pathological phenomena can be subject to inter- and intra-observer variability, leading to discrepancies in diagnosis and treatment outcomes [123, 178, 179]. Thus, the implementation of DL algorithms can aid in establishing a common framework of robust and reliable disease assessment.

In Paper 3, our aim is to explore the development of a model for the automatic staging of papillary NMIBC tumors. Staging, as explained in Section 7.2, is based on the detection of cancerous invasive areas in the lamina propria. We adopt a multi-magnification CNN input for capturing local and global information, which proved superior to single-scale models. Although the model is trained in a supervised manner, we incorporated classes for non-diagnostically relevant tissue to enable the independent differentiation of ROIs from artifacts, among other factors. For a minor trade-off in performance, we achieved a pathologist-free model, that is, a fully automated system without the need for ROI extraction at the inference stage. Nevertheless, this research sets a foundation toward automatic staging of NMIBC.

Another significant clinical risk factor is overall patient grading, explored in Paper 4. We introduce an automated pipeline, NMGrad, that leverages a tissue segmentation algorithm for ROI delineation of NMIBC WSIs, followed by NMIA from Paper 1 for predicting the ultimate WSI grade. Tile extraction was suited to NMIA, as the lesion was partitioned into scattered regions. Result predictions suggest that utilizing a greater number of magnifications as input leads to improved accuracy. In comparison to a state-of-the-art model [31], we obtained a slightly less biased model. Moreover, NMGrad better conforms to clinical expectations, identifying indicative HG regions rather than scattered patches.

We further investigated the significance of attention scores and grade predictions generated at inference. Heatmaps were consulted with patholo-

gists for qualitative analysis. Results on the annotated regions from the test set, showed NMGrad's ability to discern LG and HG regions. Higher attention scores correlated with HG areas in HG WSIs, but not for LG WSIs, where attention was spread widely. Furthermore, a strong correspondence was observed between the label of annotated ROIs and the output region predictions.

Also, in Paper 5, we explored the relevance of manual annotations by training on annotated regions but testing on automatically segmented ones, and vice versa, compared to a fully automated system. It was found that models performed better when trained on automatically generated ROIs and tested on annotated ones, suggesting benefits from the larger dataset including non-annotated data.

9.1.3 Predicting Clinical Outcomes in NMIBC Patients

This section is associated with the thesis sub-objective SO_3: "Propose a deep learning method for predicting clinical outcome from histological slides of NMIBC patient".

Research in CPATH predominantly focuses on diagnostic prediction, deliberately selecting ROIs during both model training and inference stages. However, when it comes to prognostics, weakly labeled data is commonly used, with patient-based labels defining treatment outcomes, recurrence, or disease progression [118, 180, 181]. Consequently, there is uncertainty regarding whether a specific region contains the requisite information, or if the essential information is genuinely present in the WSI [182, 183]. This underscores the challenge of accurately predicting prognostic outcomes solely based on weak labels and without comprehensive annotations of specific regions within the WSI.

Paper 5 encapsulates the entirety of prognostic applications in this thesis. We introduce a novel study focusing on predicting treatment outcomes using histopathological features, with a secondary investigation into recurrence prediction. We proposed a fully-automated pipeline that incorporates the tissue segmentation algorithm developed in Paper 2, a contrastive learning block for feature extraction, and the NMIA architecture from Paper 1. Then, we conducted a systematic evaluation for unearthing the optimal avenues for tackling prognostic prediction. The results demonstrate that among the various loss functions tested, the unsupervised contrastive loss yielded the best performance. Additionally, the analysis revealed that

models trained with mono-scale data tended to perform better compared to those trained with multi-scale data. Moreover, it was observed that features extracted from both the urothelium and lamina propria regions had the most positive impact in the predictive performance of the models. Leveraging domain expert knowledge played a pivotal role in enhancing our pipeline' succes, as our models capitalized on nuanced insights that might otherwise have been overlooked. Finally, we explored the integration of clinicopathological information into the analysis; however, this approach yielded unfavorable results.

Another prognostic investigation centered on NMGrad from Paper 4, where we correlated the informativeness of grading predictions to clinical outcome. We found that NMGrad's grading predictions exhibited a stronger correlation with disease progression compared to the manual grading performed by an expert uropathologist, which was utilized as the reference for weak patient-based labels during the NMGrad model training process. This suggests that NMGrad's grade prediction may be more predictive of NMIBC progression likelihood.

9.1.4 Limitations

A common constraint observed in the proposed methods is the absence of substantial pre-processing steps prior to data ingestion. Stain variability poses a challenge for deep learning models, which can be susceptible to laboratory biases. As a result, models trained on a single cohort may struggle to generalize to images showcasing diverse histological characteristics. An additional constraint noted was the development and evaluation of models solely on in-house cohorts, raising questions about their generalizability for wider adoption.

The study from Paper 3 was showcased as a proof-of-concept for automated NMIBC staging. However, the analysis omits flat lesions, and focuses solely on papillary lesions. Furthermore, the method hinges on individual patch predictions and lacks an aggregation mechanism to formulate a definitive WSI-level prediction.

The evaluation of NMGrad's grade as a clinical risk factor was conducted, revealing its potential significance in risk assessment. However, this assessment does not extend to staging. Additionally, there is a notable absence of an integrated framework that effectively leverages insights derived from various diagnostic models to develop prognostic models. Immunohistochemistry has not been explored as a viable option, despite its demonstrated

benefits in identifying pertinent prognostic markers. When considering prognostics in Paper 5, histopathological features were prioritized, although in routine practice, pathologists also integrate other data modalities, such as sequencing. Regarding the integration of clinicopathological information into our analysis, various multi-modal architectures could potentially optimize its utilization.

9.2 Conclusions and Future Work

This research has set the foundation for advancements in CPATH, offering enhanced diagnostic and prognostic capabilities. It has pioneered automated techniques for segmenting ROIs, streamlining analysis processes. Additionally, the study has elucidated the correlation between image features and clinicopathological concepts, providing valuable insights for further research and clinical applications in pathology. By uncovering how specific image features correspond to clinicopathological findings such as tumor grade or stage, our research offers a deeper understanding of the underlying disease processes. Through the exploration of various methods and applications across three sub-objectives (SO_1-SO_3), the primary goal of developing DL algorithms that provide clinic-relevant risk factors assessment of NMIBC patients has been achieved (O_1).

Beyond the realm of CPATH, our methods offer versatility for adoption in various domains. The nested attention-based architecture we propose is particularly suited for methods relying on weakly supervised learning, enabling nuanced feature extraction without the need for annotations. Additionally, our active learning strategy provides an efficient solution for endeavors seeking label-efficient methodologies, facilitating meaningful insights from datasets with limited labeled samples. These attributes permits our methods to be adaptable and applicable across disciplines, offering promising avenues for advancement in resource-efficient learning techniques.

Concerning future work, the proposed methods require validation using external cohorts to confirm their applicability. Furthermore, pre-processing of the color spectrum and different augmentation strategies might be beneficial for the widespread adoption of these methods. Regarding diagnostics, a prospective algorithm for staging should encompass all tumor stages and arrive at a definitive patient grade. Moreover, assessing the significance of the predictions as a statistically valid factor for risk stratification is imperative.

Regarding prognostics, exploring the integration of insights from individual diagnostic models aimed to enhance the quantitative analysis of disease risk. Additionally, incorporating omics and immunohistochemistry into prognostic models would better align with the comprehensive information accessible to pathologists for disease management decisions.

Part IV

Included Papers

Paper 1: Nested Multiple Instance Learning with Attention Mechanisms

Nested Multiple Instance Learning with Attention Mechanisms

S. Fuster[1], **T. Eftestøl**[1], **K. Engan**[1]

[1] Dept. of Electrical Engineering and Computer Science, University of Stavanger, 4021 Stavanger, Norway

Published in the Proceedings of the 21st IEEE International Conference on Machine Learning and Applications (ICMLA), 2022

https://doi.org/10.1109/ICMLA55696.2022.00038

Abstract:
Strongly supervised learning requires detailed knowledge of truth labels at instance levels, and in many machine learning applications, this is a major drawback. Multiple instance learning (MIL) is a popular weakly supervised learning method where truth labels are not available at instance level, but only at bag-of-instances level. However, sometimes the nature of the problem requires a more composite description, where a nested architecture of bag-of-bags at different levels can capture underlying relationships, like similar instances grouped together. Predicting the latent labels of instances or inner-bags might be as important as predicting the final bag-of-bags label, but is lost in a straightforward nested setting. We propose a Nested Multiple Instance with Attention (NMIA) model architecture combining the concept of nesting with attention mechanisms. We show that NMIA performs as conventional MIL in simple scenarios and can grasp a complex scenario providing insights to the latent labels at different levels.

10.1 Introduction

Multiple instance learning (MIL) is a weakly supervised learning method where several elements, called instances, are grouped as a single sample, often referred to as a bag [184]. Instances within the bag are individually unlabelled and unknown to the training algorithm. There is, however, a label representing the contents of the bag. Usually, in MIL, a bag label is defined as positive if at least one of the instances is positive, and negative if all instances are negative. We will refer to this as the MIL assumption. In MIL, feature representation of the instances are aggregated into a single representation, which is later used in a supervised learning setup with the weak label as ground truth.

Weakly supervised learning is suited for medical applications where patient-based, clinical labels are known, whereas detailed localized annotations in recorded biosignals or images often are unavailable. An example is in digital pathology, where histopathological whole slide images (WSI) are high-resolution digital files of scanned microscopic tissue sections from biopsies. WSI are of enormous dimensions; hence they are typically referred to as gigapixel images, and processing them at once is infeasible, thus the images are divided in patches. Furthermore, annotating such a large image in detail is very cumbersome and time-consuming due to the size of the image, and tumours being a heterogeneous disease challenging to diagnose [16]. Therefore, the number of annotations is limited and often pathology datasets rely exclusively on clinicopathological information, where each image patch can be considered an instance. If a tissue section is cancerous, the positive (i.e. cancerous) instances would typically be localized in one or several regions that share similar cellular features. There would typically not be single positive instances surrounded by negative ones, as this mostly would correspond to missclassified patches. A conventional weakly supervised model as MIL would not perceive this sense of location since all instances are grouped in a single bag, and features will be aggregated under one bag representation [185]. One way to overcome this is to introduce Nested MIL (NMIL), allowing to group the extracted patches from regions separately while training the model on a weak label.

Even if such a nested system can perform well, the knowledge of which individual instances or inner-bags are the most impactful will be lost with a straightforward setup of MIL. Weakly supervised methods overcome the problem of lacking individual labels overshooting the problem from instances to bag prediction. This results in a low degree of explainability

and applicability for interpreting a particular prediction, which has been one of the major focus in deep learning recently [186, 187]. Attention mechanisms can be integrated in a weakly supervised model architecture to provide a degree of explainability at instance level. Attention scores reveal how input features are weighted and can be visualized as a magnitude value of the instance significance [188–191]. Retrieving attention scores would lessen the opacity of a nested architecture, which inherently broadens the distance between instances and bag label. Besides, attention-based models are showing comparable or improved performance on their bag predictions [192].

Our proposed model architecture, Nested Multiple Instance with Attention (NMIA), overcomes the intricate relationship between instance and bag labels with nesting, while offering a high degree of interpretability using attention mechanisms. To validate the idea of finding bags with multiple positive instances, simulating a region of intestest (ROI), the MNIST dataset is used, a classical image dataset consisting of handwritten digits [162]. Moreover, to prove that nesting can perform in other domains, we make use of PCAM, an image dataset consisting of patches extracted from WSI of lymph node sections [163, 164].

10.2 Related work

Several applications have been developed that use MIL for overcoming the lack of annotated ROIs in an image [193, 194]. Chen et al. [195] train end-to-end feeding an entire WSI in a strong supervision manner, adopting a variant of MIL. To provide further intuition into the composition of the bags and the relevance that individual instances carry in the classification, Chikontwe et al. [196] propose a center loss that characterizes intra-class variations by minimizing the distance among instances from the same class. Li et al. [75] propose a dual-stream architecture to learn instance and bag classifiers at once, where the first instance would be an instance classifier and the second stream aggregates the instances into a bag embedding to feed to a bag classifier. Also, He et al. [197] use a clustering-based strategy to obtain hidden structure information in the feature space to discover positive instances. These methods work well under the MIL assumption but would not necessarily understand more *complex scenarios*, such as region-based analysis of WSI, detection of sequential events in a time-series signal, natural language processing of blocks of text, among others[174].

The concept of bag-of-bags is first seen in an application for prostate cancer detection using magnetic resonance images [198]. It is, however, Tibo et al. [173, 174], who formulate the idea of a bag within a bag. In their work, they present the multi-multi learning framework for MIL consisting of top and sub-bags; therefore, a nested implementation limited to two levels of nesting. In order to aggregate instance-level representation into a bag-level representation, they use dense layers. This implementation, however, remains opaque since its architecture does not offer interpretability for which instances or inner-bags contributed the most to the final prediction. Adding nesting exacerbates this issue, since the gap between the instances and the final weak label is even more prominent.

Attention-based MIL frameworks were first introduced by Ilse et al. [84]. Attention mechanisms give insight into the model's decision making and ability to pick out instances of interest. A trainable attention mechanism identifies the instances that have a more significant influence in making a positive prediction. This is self-enforcing by using the attention scores to strengthen the instance representation before aggregation. The final prediction is made using the aggregated representation. Attention mechanisms have proven to be an efficient method to improve the model performance and provide relevant insight into the model decision-making and meaningful data points, as many works have followed this trend since the original publication [199–203].

In this work, we propose Nested Multiple Instance with Attention (NMIA), a novel model architecture for weakly supervised learning methods on structured data. Our proposed architecture defines any number of levels; hence, a general implementation for nested MIL is presented. Furthermore, our experiments demonstrate the applicability of deeper nesting with a 3-level architecture and combining it with attention mechanisms for interpretability.

10.3 Methodology

10.3.1 Nested Multiple Instance Learning

Multiple instance learning (MIL) is a weakly supervised method trained in a supervised manner considering outermost bags and their corresponding labels. For the conventional MIL binary setting, a dataset $\mathcal{X}, \mathcal{Y} =$

$\{(\mathbf{X}^i, y^i), \forall i = 1, ..., N\}$ is formed of pairs of sample sets $\mathbf{X}$ and their corresponding labels y, where i denotes the current sample for a total of N samples. A sample $\mathbf{X}$ consists of a bag of instances $\mathbf{x}_l$:

$$\mathbf{X} = \{\mathbf{x}_l, \forall l = 1, ..., L\} \tag{10.1}$$

where L is the number of instances in the bag. In order to obtain an instance representation or embedding from input data, for example an image patch, we make use of a feature extractor. An image feature extractor $G_f : \bar{\mathbf{x}} \rightarrow \mathbf{x}$ is typically a convolutional neural network which maps an image patch $\bar{\mathbf{x}}$ into a feature embedding vector $\mathbf{x}$. Instance representations $\mathbf{x}$ from a bag $\mathbf{X}$ are aggregated to form a bag representation using an aggregation function Ξ, which can, for example, be replaced by either the *mean* or *max* operators.

A label $y \in \{0, 1\}$ is associated with the bag $\mathbf{X}$. Although each instance might be associated with a label y_l, they are generally unknown; hence only bag labels are used during training. At the inference stage, a test set might be associated with labels both at the bag and instance level to provide performance metrics. Under the conventional binary classification MIL assumption, a bag label is positive with the single presence of a positive instance. Then, the model learns comparing y with the prediction $\hat{y}$, computed by the bag classifier Θ_c.

$$y = \max_l \{y_l\} \tag{10.2}$$

$$\hat{y} = \Theta_c(\Xi(\mathbf{X})) \tag{10.3}$$

In contrast with conventional MIL, NMIL setting consists of levels of bags within bags where only the innermost bags contain instances. In Figure 10.1, the idea of grouping instances in inner-bags, inner-bags in larger bags, and finally in one outermost bag is illustrated.

Let J denote the number of nested levels, K_j number of bags at level j and $L_{j,k}$ denote the number of instances or bags in a bag k. A set $\mathbf{X}_j$ contains a set of inner-bags $\mathbf{X}_{j,k}$:

$$\mathbf{X}_j = \{\mathbf{X}_{j,k}, \forall k = 1, ..., K_j\} \tag{10.4}$$

for j defining the current nesting level up to J levels of nesting. For $J = 1$, NMIL with one level of nesting, the NMIL notation corresponds

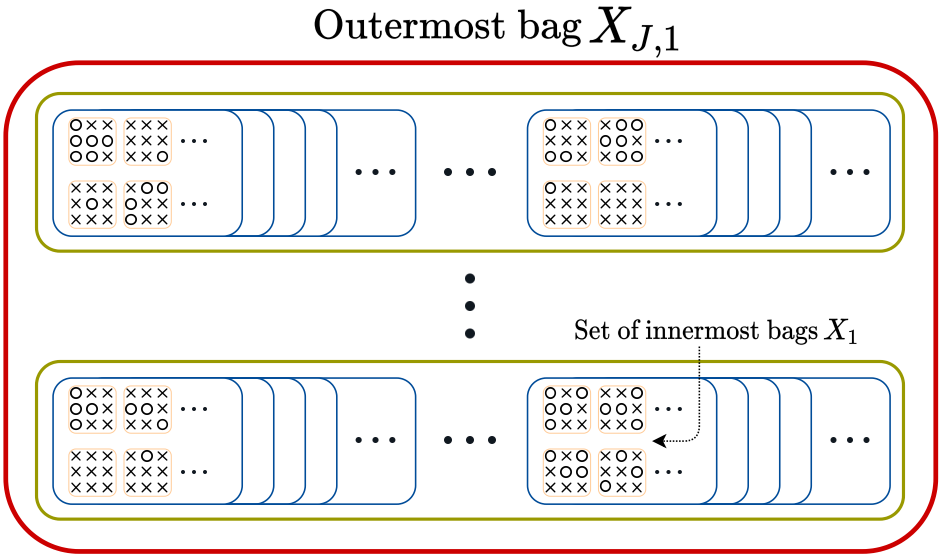


Figure 10.1: Nested bag-of-bags. Instances from two different classes (crosses and circles) are drawn into the innermost bags to form sets which are recursively grouped finally forming the outermost bag.

to the ordinary MIL notation described in Eq. (14.4). The number of inner-bags K_j can vary from level to level. For a given inner-bag k at level j, where l defines the instance number up to $L_{j,k}$, a bag of instances $\mathbf{X}_{j,k}$ is expressed as:

$$\mathbf{X}_{j,k} = \{\mathbf{x}_{j,k,l}, \forall l = 1, ..., L_{j,k}\} \tag{10.5}$$

By latent labels, we refer to the actual, but in general unknown, labels of an instance or inner-bag. At the training stage, however, these are necessary to determine the final bag-of-bags weak label y. For simplicity in the experiments section, we will refer to an instance latent label $y_{j,k,l}$ as y_l^j to describe the latent label of an instance l in a level j, omitting bag index k.

10.3.2 Attention Mechanism

According to Ilse et al [84], a multiple instance attention block is constructed using an embedding-level approach to obtain a bag-level representation from the instances within, as depicted in Figure 10.2. The proposed attention module's input corresponds to low-dimensional embeddings. These are generated by a feature extractor or by bag embedding representations of

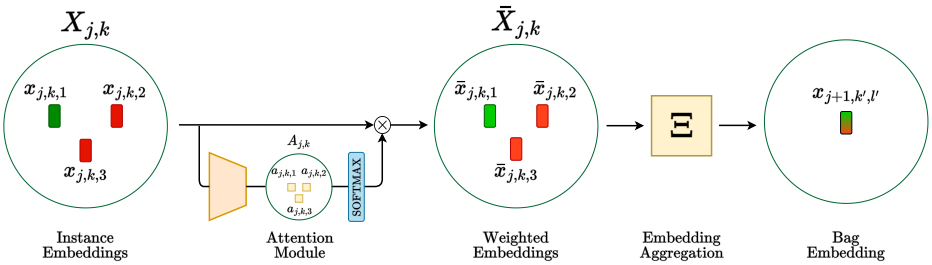


Figure 10.2: MIA block. Input embeddings are fed into the attention module to compute attention scores. Green instances represent a positive class, while a red represents a negative class. Those scores are used to compute weighted representations of instance embeddings. Then, these weighted embeddings are aggregated to create a final bag embedding.

previous levels. This module observes those embeddings and computes attention scores that leverage the meaningfulness of the features extracted for the given task. Finally, the attention scores are aggregated to obtain the bag representation.

For ease of exposition, we define indexes $\delta = (j, k, l)$, $\delta' = (j+1, k', l')$, $\gamma = (j, k)$. We omit the use of training sample index i. An attention score a_δ for a given input embedding x_δ is calculated as:

$$a_\delta = \exp\{\mathbf{w}^\top(\tanh(\mathbf{V}\mathbf{x}_\delta^\top) \odot \text{sigm}(\mathbf{U}\mathbf{x}_\delta^\top))\} \tag{10.6}$$

where $\mathbf{w} \in \mathbb{R}^{L\times 1}$, $\mathbf{V} \in \mathbb{R}^{L\times M}$ and $\mathbf{U} \in \mathbb{R}^{L\times M}$ are trainable parameters and $\odot$ is an element-wise multiplication. Furthermore, the hyperbolic tangent $\tanh(\cdot)$ and sigmoid $\text{sigm}(\cdot)$ are included to introduce non-linearity for learning complex applications. Then, attention scores a_δ are normalized into $\tilde{a}_\delta$ to ensure that the sum of the components of the attention scores vector is 1, as this makes it possible to have variable bag sizes. Note that the unnormalized a_δ would better reflect the attention score directly, and we use that for visualization in the experiments.

$$\tilde{a}_\delta = \frac{a_\delta}{\sum_{l=1}^{L_\gamma} a_\delta} \tag{10.7}$$

Finally, the aggregation function Ξ transforms a leveraged bag of embeddings $\tilde{\mathbf{X}}_\gamma$ to obtain a bag representation $\mathbf{x}_{\delta'}$ as:

$$\mathbf{x}_{\delta'} = \Xi(\tilde{\mathbf{X}}_\gamma) = \Xi(\{\tilde{a}_\delta \cdot \mathbf{x}_\delta, \forall l = 1, ..., L_\gamma\}) \tag{10.8}$$

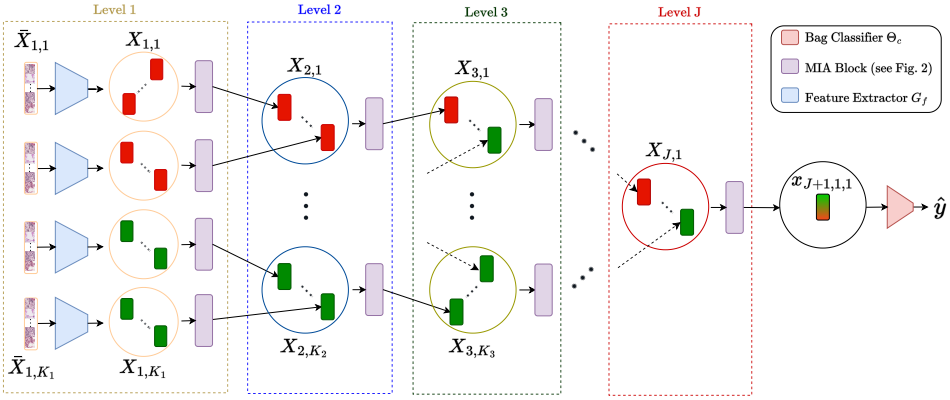


Figure 10.3: NMIA model architecture. The feature extractor G_f projects all instances into low-dimensional embeddings. Consecutive MIA blocks aggregate deeper levels into more superficial representations. Finally, a bag-of-bags embedding is fed to the classifier Θ_c for obtaining a bag prediction $\hat{y}$.

10.3.3 Model Architecture

The proposed neural network architecture NMIA combines a recursive bag processing of low-dimensional embeddings and attention mechanisms. The multiple instance with attention (MIA) block computes attention scores for each input embedding and aggregates them to create an embedding that represents the bag's contents, see Figure 10.2. Instances are transformed into embeddings, weighted and aggregated into a bag-of-bags embedding after passing through the MIA blocks, see Figure 10.3. As a result, a bag-of-bags embedding $\mathbf{x}_{J+1,1,1}$ is formed. Finally, a classifier Θ_c predicts the label $\hat{y}$ for the sample $\mathbf{X}$. The model's forward propagation is summarized in Algorithm 3. Although the example given corresponds to features extracted from images, this problem can be extrapolated to other signal types that can be represented as feature embeddings.

10.4 Experimental Setup

Several experiments were carried out to show the usefulness of NMIA compared to multiple instance (MI) architectures under different types of data and tasks. Also, we show the importance of attention-based models to obtain further interpretability from meaningful instances and compare the performance to traditional aggregation techniques. Two image datasets were used: MNIST [162] and PCAM [163, 164]. MNIST consists of a training

Algorithm 3 NMIA Forward Propagation

Input: image bag-of-bags sample $\bar{\mathbf{X}} = \{\bar{\mathbf{x}}_l, \forall l = 1, ..., L\}$
Output: classifier prediction $\hat{y}$

```
/*Transform input image instances into feature embeddings*/
for l = 1 to L do
    x̄_l = G_f(x_l)
end for
for j = 1 to J do
    for k = 1 to K_j do
        for l = 1 to L_{k,j} do
            /*Define indexes δ = (j,k,l), δ' = (j+1,k',l')*/
            a_δ = exp{w^⊤(tanh(V x_δ^⊤) ⊙ sigm(U x_δ^⊤))}
        end for
        ã_δ = a_δ / Σ_{l=1}^{L_{k,j}} a_δ
        x_δ' = Σ_{l=1}^{L_{k,j}} x̃_δ = Σ_{l=1}^{L_{k,j}} x_δ ã_δ
    end for
end for
/*Resulting bag-of-bags embedding is fed into the classifier*/
ŷ = Θ_c(x_{J+1,1,1})
return ŷ
```

and test set of 60,000 and 10,000 examples, respectively. PCAM consists of 327,680 patches, where each patch is annotated in a binary manner to indicate the presence of metastatic tissue. All models were trained using stochastic gradient descent (SGD) optimizer and early stopping. Binary cross-entropy loss function is applied over the resulting prediction from the classifier and the given weak label, while the attention mechanisms remains unconstrained. The attention scores are used for interpretability proof.

A custom convolutional neural network and VGG16 were used as feature extractors G_f for MNIST and PCAM, respectively. The models were trained and tested with 20,000 and 5,000 bag-of-bags, respectively, constructed by randomly extracting instances within the class of desired instance labels. All models are implemented in Python 3.6 using Tensorflow machine learning library [204][1].

We have conducted three experiments as follows. In the first two, we have considered the positive instance class $y_{l^+}^1$ as the digit 9 for MNIST

[1] github.com/Biomedical-Data-Analysis-Laboratory/NMIL_Attention

and tiles with metastatic tissue for PCAM. Dataset samples were arranged to form a 2-level setting. The third experiment was carried out exclusively for MNIST in a 3-level setting. Latent labels y_l^j introduced in intermediate levels are formulated only to obtain the resulting bag-of-bags labels, and to be compared with attention scores to evaluate if we find the correct latent labels. All models are trained entirely on weak bag-of-bags labels y.

Exp1: A dataset is constructed following the MIL assumption described in Eq. (10.2). In this assumption, there is nothing to gain in using NMIA, but we want to show that the NMIA architecture is flexible and the model will perform comparably to MI models. A conventional MI model both with and without attention is compared to 2-level NMI models with random grouping of the bags at the first level.

Exp2: A dataset is constructed such that at least two instances from the same inner-bag have to be positive for the weak bag-of-bags label to be positive, as described in Eq. (10.9). This is motivated from region-based analysis of medical images, where typically an object belonging to a positive class is located in a specific region and not scattered across the entire image. Therefore, this particular positive region of the image will contain several positive instances. Regions containing few positives are regarded as noise or misclassified instances, and they should not be reflected in the overall prediction.

$$y_l^2 = \begin{cases} 0, & \#y_{l+}^1 \leq 1 \\ 1, & \#y_{l+}^1 > 1 \end{cases} \quad \overset{\mathbf{1^{st}}}{} \qquad y = \begin{cases} 0, & y_l^2 \notin 1 \\ 1, & y_l^2 \in 1 \end{cases} \quad \overset{\mathbf{2^{nd}}}{} \tag{10.9}$$

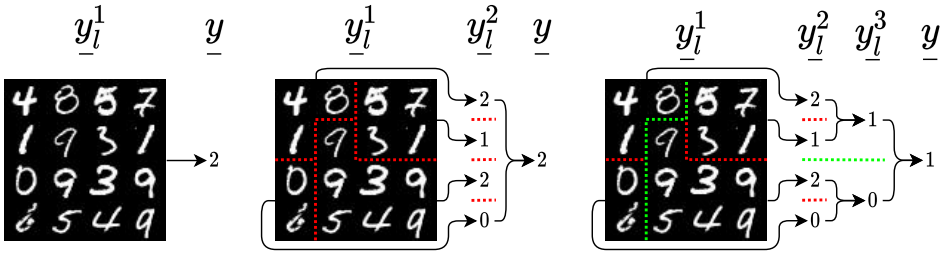


Figure 10.4: From left to right, 1, 2 and 3-level partitioning. Red dotted lines separate bags-of-instances at a second level, while greens at a third level. The task of *Exp3* is to find out if a second-level bags contains at least one first-level bag with odd numbers but no first-level bags containing only even numbers.

Exp3: Here, we want to find a ROI that contains a bag of odd numbers and not one of even numbers and a 2-level solution is not enough to overcome

this task; hence three levels of nesting are required. A region in the image is considered 0 if all instances are even numbers, 1 if they are odd numbers and 2 if there is a mix. In Figure 10.4, an example is shown where the entire image is one region, the image partitioned in regions and regions with sub-regions, respectively. A 3-level partitioning can be used to construct a dataset reflecting a complex scenario as described in Eq. (10.10). To get a final bag label to 1, there has to be second level region that contains a bag of only odd numbers but not any bags of only even numbers. Such outline is impossible to learn using MI models, as proven in *Exp2*; thus, corresponding MI and MIA tests were never carried out.

$$
\begin{aligned}
& \mathbf{1^{st}} && \mathbf{2^{nd}} \\
y_l^2 = &\begin{cases} 0, & y_l^1 \in \{0,2,4,6,8\} \\ 1, & y_l^1 \in \{1,3,5,7,9\} \\ 2, & otherwise \end{cases} && y_l^3 = \begin{cases} 0, & y_l^2 \in \{0 \cap \bar{1}\} \\ 1, & y_l^2 \in \{\bar{0} \cap 1\} \\ 2, & otherwise \end{cases} \\
& && \mathbf{3^{rd}} \\
& && y = \begin{cases} 0, & y_l^3 \notin 1 \\ 1, & y_l^3 \in 1 \end{cases}
\end{aligned}
\tag{10.10}
$$

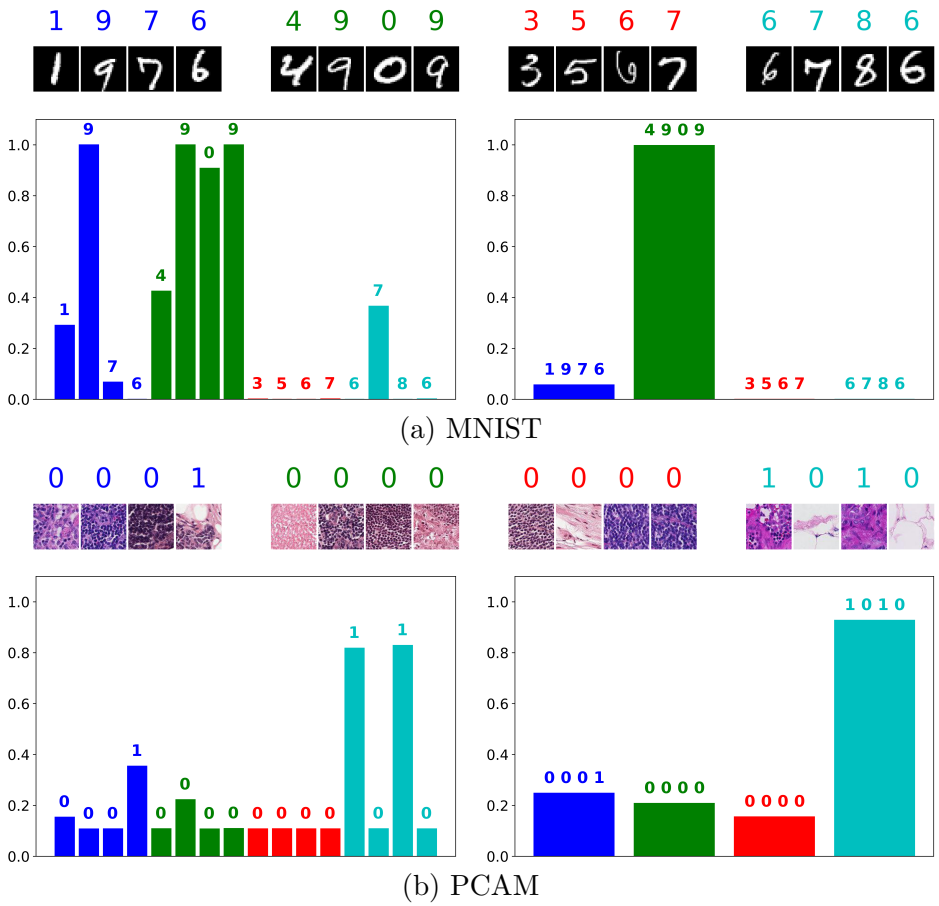
10.5 Results & Discussion

F1 scores for the experiments are listed in Table 10.1. MI architecture is compared to MI with attention (MIA), nested MI architecture (NMI) and nested multiple instance architecture with attention, the proposed NMIA architecture. Note the absolute value of the F1 score is dependent on the chosen feature extractor which is not the focus of this paper, rather the relative values between the MI, MIA, NMI an NMIA in simple scenarios (*Exp1*) and more complex scenarios (*Exp2*, *Exp3*) is what we seek to demonstrate.

Exp1 presents a setup where individual instance latent labels are directly responsible for the resulting bag weak label. Here, we can see that a conventional MI model can perform highly and the choice of architecture does not affect the predictive power of the classifier. Nesting becomes irrelevant when only the individual labels of the instances are meaningful, but not their arrangement across inner-bags.

However, for *Exp2*, we show that a conventional MI architecture breaks down because it cannot perform or even understand the nature of the task.

From the MI model perspective, all instances are at the same level and belong to the same set. *Exp2* and *Exp3* were designed to demonstrate the strength of nesting when the relationship among instances and inner-bags is fundamental for obtaining the final weak label. NMIA can process these subsets independently, hence understanding the relationship among instances.



(a) MNIST

(b) PCAM

Figure 10.5: Examples of attention scores a_δ in test samples from *Exp2* with 2-level NMIA for (a) MNIST and (b) PCAM. A single positive instance in the inner-bag is considered noise whereas two or more should give a positive inner-bag, resulting in a positive bag-of-bags bag. Instances are categorized by colors indicating inner-bag belonging, while digits represent the true label of the instance. Bar plots on the left column show the attention at level 1, while on the right, at level 2. Positive instances obtain the highest attention scores.

Table 10.1: F1 scores for experiments on MNIST and PCAM datasets.

	MNIST			PCAM	
	Exp1	***Exp2***	***Exp3***	***Exp1***	***Exp2***
MI	0.929	0.345	N/A	0.957	0.290
MIA	0.957	0.472	N/A	0.973	0.286
NMI	0.923	0.855	0.556	0.964	0.700
NMIA	**0.959**	**0.921**	**0.836**	**0.978**	**0.734**

Furthermore, implementing a model with an attention mechanism provides an edge over the model that does not, both on performance and interpretability, as shown in Figure 10.5. We can observe that the attention mechanism can correctly identify the positive instances at the instance level and recognise the positive inner-bags, distinguishing them from those containing noisy instances. A conventional weakly supervised model would not perceive this sense of location since all instances are encapsulated under the same bag. This results in more interpretable predictions on account of which instances and inner-bags the model weighs more.

Exp3 further demonstrates NMIAs strengths in a 3-level setup. As mentioned in the definition of the experiment, MI and MIA were not tested out due to findings explored in the previous *Exp2*. As for the nested architectures, we prove that a nested implementation efficiently handles complex scenarios, being able to achieve remarkable performance. Moreover, we notice a significant improvement in predictive power of NMIA over NMI. NMIA model architecture reaches a F1 score of 0.836, thus incorporating attention mechanisms provides a substantial boost in predictive power, as well as intelligible predictions.

10.6 Conclusions

In this paper, we have proposed the NMIA architecture for solving applications that simpler MI architectures cannot, for when dependencies among sets of bags are to be considered. We have formulated a notation for a generalized implementation of nesting for any number of levels. We have presented experiments processing interdependent subsets from images

demonstrating the flexibility and improved performance relative to MI. Moreover, NMIA can be used in a wide range of applications due to its flexibility. Finally, implementing an attention mechanism helps identify key instances and inner-bags contained in a set of bags, giving insight into the practical relationship between latent labels and the attention given among different levels. Future research will consider the potential effects of NMIA with attention on a full-scale medical imaging dataset more carefully.

Paper 2:
Active Learning Based Domain Adaptation for Tissue Segmentation of Histopathological Images

Active Learning Based Domain Adaptation for Tissue Segmentation of Histopathological Images

S. Fuster[1], F. Khoraminia[2], T. Eftestøl[1], T. C. M. Zuiverloon[2], K. Engan[1]

[1] Dept. of Electrical Engineering and Computer Science, University of Stavanger, 4021 Stavanger, Norway

[2] Dept. of Urology, Erasmus MC Cancer Institute, University Medical Center, 3015 GD Rotterdam, The Netherlands

Published in the Proceedings of the 31st European Signal Processing Conference (EUSIPCO), 2023

https://doi.org/10.23919/EUSIPCO58844.2023.10290058

Abstract:
Accurate segmentation of tissue in histopathological images can be very beneficial for defining regions of interest (ROI) for streamline of diagnostic and prognostic tasks. Still, adapting to different domains is essential for histopathology image analysis, as the visual characteristics of tissues can vary significantly across datasets. Yet, acquiring sufficient annotated data in the medical domain is cumbersome and time-consuming. The labeling effort can be significantly reduced by leveraging active learning, which enables the selective annotation of the most informative samples. Our proposed method allows for fine-tuning a pre-trained deep neural network using a small set of labeled data from the target domain, while also actively selecting the most informative samples to label next. We demonstrate that our approach performs with significantly fewer labeled samples compared to traditional supervised learning approaches for similar F1-scores, using barely a 59% of the training set. We also investigate the distribution of class balance to establish annotation guidelines.

12.1 Introduction

Computer-aided diagnosis (CAD) systems that utilize machine learning techniques for medical imaging analysis have been shown to be an effective way to reduce subjectivity and speed up the diagnostic process [123]. Digital microscopy scanners are capable of generating high-resolution digital images from scanned tissue sections, also known as Whole Slide Images (WSI). These images can are pre-stored at various magnification levels, allowing pathologists to adjust the zoom level like they would with physical microscopes. Lower magnification is typically used to view tissue-level morphology, while higher magnification is useful for examining cell-level features.

Bladder cancer is among the most commonly diagnosed cancer types. According to the World Health Organization, over 573,000 new cases and 213,000 deaths were registered in 2020 [205]. WSIs from bladder cancer are highly disorganized scanned tissue sections for several reasons. Urothelial carcinomas often present papillary structures, elongated finger-like bundles of tissue, that alter the normal appearance of the urothelial lining. Also, transurethral resection of a bladder tumour (TURBT) is a complicated operation that difficults a clean tumour extraction as the cauterization process leaves damaged tissue areas. As a result, a significant amount of artifacts and other non-diagnostically-relevant tissue is present within the slides. Using all the regions of a WSI for model training or inference as input would often add unnecessary noise. Manual annotations of potential regions of interests (ROIs) is an arduous, time-consuming and labor-intensive task. Hence automatic methods have emerged to reduce this time constraint. Tissue segmentation in computational pathology (CPATH) enables the analysis of specific ROIs within a WSI and can also improve the performance of the model by actively selecting the most informative tissue types [206, 207]. A Multiscale Approach for Whole-Slide Image Segmentation of five Tissue Classes in Urothelial Carcinoma Slides [26] proposes a multiscale convolutional neural network (CNN) that can effectively segment five different tissue classes in non-muscle invasive bladder urothelial carcinoma slides. The model classifies all input areas into blood, damaged, muscle, stroma and urothelium tissue. They demonstrate that their approach outperforms existing methods and is able to handle the large size and variability of WSIs within their private cohort. The presence of a domain shift between images obtained from different laboratories impedes the performance of deep learning models on out-of-distribution samples

[208]. Thus, implementing the algorithm on a new dataset may require of an adaptation. However, the cost of labeling resources is critical, especially in the clinic.

When dealing with limited data, it is frequently necessary to resort to the implementation of deep learning methodologies that are more cost-efficient, such as active learning (AL). AL is a variant of supervised learning (SL) involving human interaction during training, also referred to as having a human in the loop during training [209]. The goal of an AL setup is to actively interfere during the training procedure, extract new data points from the classes the model struggles to comprehend and append them to the training set [210]. Therefore, an human in the loop is expected to spend time resources for labeling data points from the class the model demands. Recent works within the field of CPATH have demonstrated implementing AL techniques results in similar performances to SL, with constrained data settings [166–172]. Analogously, AL can enhance the tissue segmentation model from [26] by selecting informative samples from an undetermined dataset with limited annotations. Despite the publication of protocols for critical bias assessment of clinical models, no official guidelines have been established regarding the required numbers of annotations, images, and laboratories to capture the variation present in real-world data [211, 212]. The need for more well-defined class sampling strategies within the field of histopathology arises. Due to privacy concerns, medical data cohorts often cannot be made publicly available. This leads to limited model predictive generalization as medical applications are developed for a target dataset. Consequently, domain adaptation of deep learning models into unexplored data domains is needed.

In this work, we propose a domain adaptation framework for deep learning models within the field of histopathology. We choose to adapt an algorithm that segments bladder cancer WSIs into different tissue types. The model architecture uses a multiscale CNN backbone that incorporates information from different magnification levels, which has been developed using another dataset from a different hospital. In order to adapt the model to a new unseen domain, we adopt an AL strategy for a more efficient labeling effort. We proactively select samples to be included into the training data based on preemptive results on the validation set. We show that the proposed AL approach is more profitable and can be integrated to reduce labeling costs. On top of that, we also aim to guide pathologists in which annotations to provide deep learning models before investing substantial amounts of effort. We estimate a balance between class distribution and model performance,

validated using a small initial subset of annotations. By assessing the model's initial performance, pathologists can then proceed with further annotations guided by this intuition.

12.2 Material and Methods

12.2.1 Dataset

We have collected a set of high-risk non-muscle invasive bladder cancer (HR-NMIBC) WSI from the first TURBT from a multi-centre cohort provided by Erasmus MC, Rotterdam, The Netherlands. WSIs were stained with Haematoxylin and Eosin (H&E) and scanned using a 3DHistech P1000 scanner at 80x magnification stored as MRXS files. The total number of slides is 155, for which a pathologist has annotated the slides with tissue types. Data heterogeneity produces more generalizing models than using higher amounts of data from the same slides [169]. Thus, a pathologist was asked to annotate areas in a rough, imprecise manner in order to obtain several annotated regions per WSI, within a time limit. The time usage was limited to a maximum of one hour per WSI, including diagnostic labels not used in this work. As a result, some regions were annotated in every WSI, but not the entirety of present tissue was annotated. The aim of this annotation protocol was to collect diverse scenarios, hence capture the tissue heterogeneity characteristic from bladder cancer. Moreover, to avoid incorporating a human in the loop during training, we preemptively collected available tissue type annotations and defined pools of data to draw from. In total, 127, 16, and 12 WSIs were annotated and used for training, validation and test, respectively, where the split is done on WSI level to avoid cross-contamination. Tiles from annotated regions of urothelium, stroma, muscle, blood and damaged tissue were extracted from the EMC cohort, using the strategy proposed in [26] where more details can be found. In short; tiles were extracted at 2.5x, 10x and 40x magnifications, thus forming a triplet, and a maximum of 500 triplets were extracted per WSI. A tile size of 128×128 was used, and all three tiles in a triplet share a common physical point as the center pixel. Therefore, tiles at a lower magnification cover a larger physical area. The total number of tiles across sets is stated in Table 12.1.

	Train	Val	Test
WSI	127	16	12
Blood	35105	2500	4106
Damaged	65920	2500	57296
Muscle	67007	2500	50347
Stroma	86978	2500	79338
Urothelium	91006	2500	38813
Total	346016	12500	229900

Table 12.1: Number of WSI and tiles per tissue type (class) in train/val/test split of the available dataset of high-risk non-muscle invasive bladder cancer.

12.2.2 Model Architecture

The tissue segmentation model TRI-CNN proposed in [26] is adopted into our pipeline. This model was trained using WSIs of NMIBC patients from Stavanger University Hospital (SUS), Stavanger, Norway. The model architecture consists of a multiscale CNN setup that aggregates local and global information. Triplets are fed through three weight-independent VGG16 backbones trained for each of the magnifications, concatenating each output feature vectors before classification. Then, the formed feature vector is fed through the classifier to predict a tissue class, using softmax activation. A representation of the model architecture is presented in Fig. 11.2.

12.2.3 Active Learning Procedure

Active learning (AL) can be a powerful tool for improving the performance of deep learning models when labeled data is limited [210]. In AL, the model is initially trained on a small labeled dataset. Thereafter, based on intermediate performance metrics and a query strategy for requesting additional samples of the most informative class, a human annotator is asked to label a small number of additional samples. The model then uses this newly labeled data to update its parameters and improve its performance. This process is repeated until either the model reaches a satisfactory level of performance, the resources are exhausted or a set number of iterations has been conducted. One of the main advantages of AL is that it can significantly reduce the amount of labeled data required to train a deep

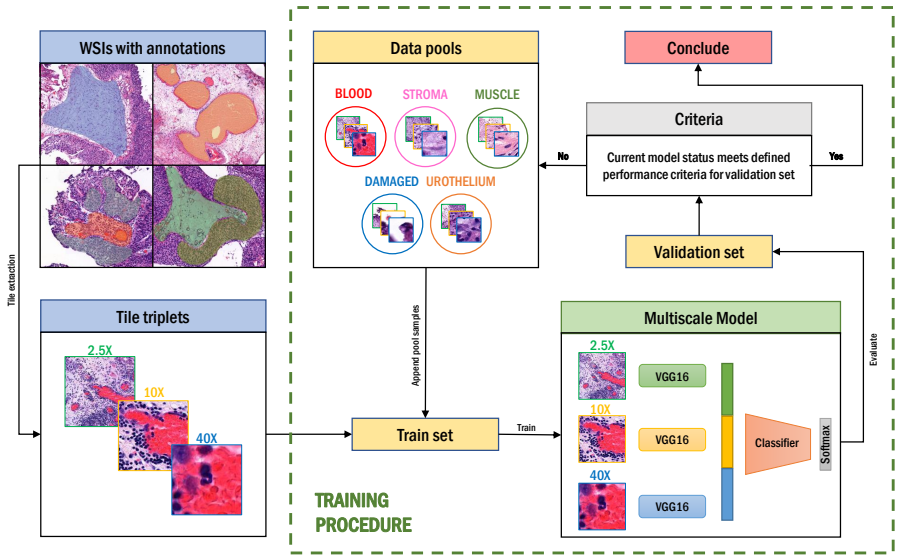


Figure 12.1: Active learning framework TRI-AL. Pathologists annotate ROIs, from which tile triples are extracted at different magnifications levels (2.5x, 10x, 40x) to form a triplet. During the training stage, an starting training set is defined for training a multiscale model. The model performance is evaluated on the validation set. Then, stopping criteria decides to resume with another training iteration or conclude the learning. In case that the criteria is not met, new samples from the pools of data are drawn and appended to the current version of the train set.

learning model to achieve acceptable performance. Additionally, active learning can also improve the model's performance by allowing it to adapt to changing conditions or concepts over time. However, AL also has some limitations. It requires human intervention, which can be time-consuming or cumbersome to organize, and it may introduce bias into the training process. Overall, the choice between SL and AL will depend on the specific needs of the application and the availability of labeled data.

AL strategies, or query strategies, involve analyzing the information value of unlabeled instances. The most commonly used query strategy is uncertainty sampling [209, 213]. This framework revolves around the model uncertainty in labeling certain instances, revolving in probabilistic methods, such as entropy and prediction confidence. Predictions on unlabeled instances presenting a high degree of entropy are selected for inclusion to the training data. These are later transferred to a domain expert for labeling. Inspired by these methodologies, but with the goal

of domain adaptation, our sampling strategy falls under the uncertainty query strategies. For the AL procedure, we further divided the training set, $\mathcal{X}$, into initial training subset $\mathcal{X}_0$ and pools of training data $\mathcal{P}$. The algorithm selects new samples based on model performance. Considering the false negative ratio (FNR) on the validation set $\mathcal{V}$ and total sample size S, N_j^i samples from the pools $\mathcal{P}$ are appended into the current training set, for every class i on iteration j. Furthermore, we define a relative class balance size δ_i^j to represent the percentage of data points for class i on the current training set $\mathcal{X}_j$. These steps are iteratively repeated until $\mathcal{P}$ cannot provide more samples of the requested class or J iterations have passed, see Algorithm 4 for more details.

Algorithm 4 Active Learning Train Procedure

procedure $ALtrain(\mathcal{X}, \mathcal{V}, J, S)$
$\mathcal{X}_0, \mathcal{P} \leftarrow \mathcal{X}$ ▷ *Split train set into initial subset and pools*
while $j < J$ **or** $len(\mathcal{P}_i) > N_j^i$ **do**
  for $j \leftarrow 1$ to J **do**
    $MODEL_{TRAIN}(\mathcal{X}_j)$
    $MODEL_{EVAL}(\mathcal{V})$
    for $i \leftarrow 1$ to I **do**
      $N_j^i = \frac{FNR_{j-1}^i \cdot S}{\sum_i FNR_{j-1}^i}$ ▷ *Per class new data points*
      $\mathcal{X}_j^i \xleftarrow{+} sample(\mathcal{P}, N_j^i)$
    end for
  end for
end while
return $\hat{y}$

12.3 Experiments

All experiments are done using transfer learning from TRI-CNN, using the same hyperparameters as in the original algorithm [26]. All models were trained using unfrozen VGG16s, a multi-class cross entropy loss function, learning rate of 1.5e-4 and SGD optimizer. Early stopping was enabled for interrupting the training if there was no improvement in the validation loss for five consecutive epochs. The models were trained five times in order to determine the mean and standard deviation of the results.

We evaluate the performance of TRI-CNN on our test set; then, we compare it to the proposed SL and AL strategies using data from Erasmus MC, Rotterdam, The Netherlands. For our experiments, we compare the performance of standard supervised learning (TRI-SL) on different sampling of the training set, in order to leverage the performance at various thresholds of dataset sizes. Based on the entire available training set, we trained models using a fraction of the data, namely from 20 to 100%, in intervals of 20%. As for the AL methods, we define a stopping criteria based on resource exhaustion and a fixed number of iterations. The initial training set $\mathcal{X}_0$ is made of 25000 triplets per class. Every iteration i, 20000 new samples S are added to training data $\mathcal{X}_j$. We defined a maximum of 5 iterations for training the model TRI-AL$_{\text{ITER}}$, while the model TRI-AL$_{\text{OOD}}$ refers to the model trained until a class pool runs out of data points. For comparison matters, we trained an algorithm TRI-AL$_{\text{ENT}}$ based on entropy uncertainty. In this case, 30000 triplets from the pools $\mathcal{P}$ are sampled based on class distribution δ of the entire training set $\mathcal{X}$. From these, the top 20000 triplets with the highest entropy are selected. The code can be found at GitHub[1].

12.4 Results & Discussion

Results in terms of F1 scores are presented in Table 12.2. First, we observe that TRI-CNN, although achieves a respectable performance, falls short in comparison to the models trained on the new cohort, thus showing that domain adaptation is needed. TRI-CNN achieved a micro F1 score of 73.06. Regarding TRI-SL models, the performance improved along with the size of the dataset, considering that the best SL model is that using the entire training set. Nevertheless, AL strategies achieve higher performance using data more efficiently. The query strategy in TRI-AL$_{\text{ITER}}$ reaches a micro average F1 score of 90.34, compared to 90.23 for SL using 100% of the training samples. The AL data accounts for 59% of the available training samples. It is also worth mentioning the disparity in terms of performance for the blood tissue class in regards to other tissue types. One of the reasons refers to regular missclassifications. Most of the blood annotations are those of vessels present within the stroma, so its only natural that blood can be misinterpreted as stroma. However, the main reason is that the number of blood samples in the test set for blood is heavily limited in

[1] http://bit.ly/3J5XRiz

comparison to the other classes. Naturally, a minor number of blood false positives results in a significant downgrade in terms of metrics. Focusing on the TRI-AL$_{\text{ITER}}$ model performance results, blood has a recall of 86.31 while for stroma is 86.77, barely a 0.66 difference.

The development of the training set class distribution for TRI-AL$_{\text{ITER}}$ in comparison to the ones used for TRI-SL can be seen as described in Fig 12.2. Looking at the relative data balance per class, we identify a tendency over the iterations where damaged, muscle and urothelium class distribution δ reach an asymptote. Also, we observe that the algorithm deems to suggest that stroma is significantly harder to learn and requires a more extensive number of data points in comparison to other classes. Even then, the margin for δ progressively slows down and we would expect it to become stagnant.

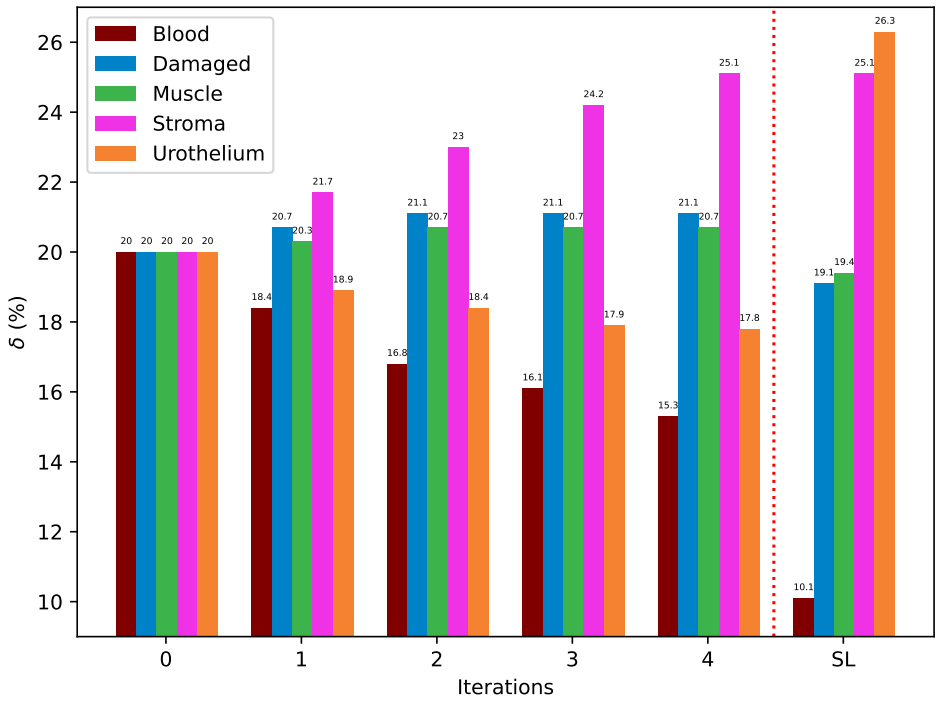


Figure 12.2: Class relative size per class δ_i over the iterations for TRI-AL$_{\text{ITER}}$ model. We observe that the algorithm prioritizes harder classes over simpler ones as the iterations pass. As $\mathcal{X}_j$ increases, δ_i reaches an asymptote for damaged, muscle and urothelium tissue classes, while penalizing blood in favor of stroma. Class distribution does not match that annotated from a pathologist, as per the TRI-SL section.

	Original [26]	TRI-SL_{20}	TRI-SL_{40}	TRI-SL_{60}	TRI-SL_{80}	TRI-SL_{100}	**TRI-AL_{ITER}**	**TRI-AL_{OOD}**	**TRI-AL_{ENT}**
Blood	29.06(-)	53.05(2.33)	**58.46(2.50)**	55.21(3.24)	53.31(3.06)	56.22(2.34)	55.36(3.50)	54.44(2.15)	51.84(2.94)
Damaged	75.65(-)	89.28(0.39)	91.82(0.23)	**92.77(0.27)**	91.55(0.66)	92.44(0.31)	92.65(0.22)	92.12(0.17)	92.63(0.31)
Muscle	71.54(-)	89.67(0.49)	89.37(0.56)	92.11(0.71)	92.51(0.64)	92.04(0.70)	**92.69(1.15)**	92.37(0.35)	92.11(0.57)
Stroma	70.73(-)	87.59(0.29)	86.93(0.76)	88.31(0.92)	88.03(1.06)	**88.47(0.64)**	88.40(1.24)	88.07(0.49)	87.41(0.58)
Urothelium	82.45(-)	90.48(0.75)	91.89(0.18)	92.66(0.30)	92.20(0.64)	93.41(0.28)	**93.45(0.46)**	93.19(0.09)	93.39(0.27)
Total (micro)	73.06(-)	88.01(0.30)	88.92(0.54)	90.08(0.41)	89.64(0.53)	90.23(0.30)	**90.34(0.66)**	89.95(0.23)	89.69(0.41)
Total (macro)	65.89(-)	82.01(0.63)	83.70(0.71)	84.21(0.68)	83.52(0.78)	**84.52(0.46)**	84.51(0.97)	84.04(0.42)	83.48(0.74)

Table 12.2: F1 scores per tissue type. We compare the model pretrained TRI-CNN on another dataset, to supervised learning (TRI-SL) approaches with varying training data sizes, to entropy-based active learning (TRI-AL_{ENT}) and our proposed active learning models (TRI-AL_{ITER}, TRI-AL_{OOD}).

Figure 12.3: Segmentation prediction of a WSI using the active learning model at inference stage. Areas from all tissue types are extracted for analysis at the inference stage, resulting in a labelled colormap based on tile predictions. The models trained in the new cohort can discern different tissue types accurately.

Regarding the comparison of the iteration TRI-AL$_{\text{ITER}}$ and data exhaustion TRI-AL$_{\text{OOD}}$ query strategies, we noticed that allowing the algorithm to add samples until a class pool $\mathcal{P}_i$ is exhausted does not translate into overall better performance, although that might be different in case that the pool was larger. To further evaluate the performance of our model TRI-AL$_{\text{ITER}}$, we compare it to entropy-based class instance selection strategies TRI-AL$_{\text{ENT}}$. Our results indicate that, on average, TRI-AL$_{\text{ITER}}$ outperforms the entropy-based approach TRI-AL$_{\text{ENT}}$, both in terms of per-class and total aggregated metrics, indicating that it is better at distinguishing between different tissue types. However, it should be noted that we also observed lower variation in the performance of the entropy-based models between runs, as indicated by the standard deviation metrics. This suggests that the entropy-based approach may be more consistent and less susceptible to fluctuations in performance across different training iterations. Nevertheless, it is important to note that the entropy-based query strategy is more computationally expensive, as it requires the calculation of sampled data points. In contrast, our proposed method does not require any additional computation beyond the training of the model.

Pathologists also visually inspected the model's segmentation results by overlaying the predicted masks over the raw WSIs. Illustrative examples of the model's segmentation results for a representative WSI of the test set are shown in Fig. 12.3 for a ROI and the entire WSI, respectively. Upon visual inspection of the results by experts, according to TRI-CNN segmentation, we have identified four main observations: staining effects leading to false positives of blood in regions with high levels of eosin stain, non-cauterized damaged areas such as blur or folding, the risk of misinterpreting infiltrative immune cells as urothelial cells, and the potential for the model to predict urothelium with significant cytoplasm as stroma. As per the models trained in the new cohort (TRI-SL, TRI-AL), it was confirmed that these models accurately segmented and classified different tissue types.

12.5 Conclusion & Future Work

In this work, we proposed a active learning framework with a multiscale CNN for domain adaptation of a tissue segmentation model of bladder cancer histopathological images. Our proposed method achieved a F1 score of 90.34 using 59% of the training data, and outperformed supervised learning strategies that used all available samples. Our results suggest

that active learning can be an effective strategy for reducing the labeling effort in histopathological image analysis. Regarding domain adaptation, we observed that we were able to customize the model to the new domain using a small labeled set. Moreover, we also presented a suggested data annotation per class burden for tissue segmentation of bladder cancer WSI. Furthermore, this model can be introduced as a pre-processing step for other applications that require tissue segmentation for ROI extraction.

Paper 3: Invasive Cancerous Area Detection in Non-Muscle Invasive Bladder Cancer Whole Slide Images

Invasive Cancerous Area Detection in Non-Muscle Invasive Bladder Cancer Whole Slide Images

S. Fuster[1], F. Khoraminia[2], U. Kiraz[3,4], N. Kanwal[1], V. Kvikstad[3,4], T. Eftestøl[1], T. C. M. Zuiverloon[2], E. A. M. Janssen[3,4], K. Engan[1]

[1] Dept. of Electrical Engineering and Computer Science, University of Stavanger, 4021 Stavanger, Norway

[2] Dept. of Urology, Erasmus MC Cancer Institute, University Medical Center, 3015 GD Rotterdam, The Netherlands

[3] Dept. of Pathology, Stavanger University Hospital, 4011 Stavanger, Norway

[4] Dept. of Chemistry, Bioscience and Environmental Engineering, University of Stavanger, 4021 Stavanger, Norway

Published in the Proceedings of the IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), 2022

https://doi.org/10.1109/IVMSP54334.2022.9816352

Abstract:
Bladder cancer patients' stratification into risk groups relies on grade, stage and clinical factors. For non-muscle invasive bladder cancer, T1 tumours that invade the subepithelial tissue are high-risk lesions with a high probability to progress into an aggressive muscle-invasive disease. Detecting invasive cancerous areas is the main factor for dictating the treatment strategy for the patient. However, defining invasion is often subject to intra/interobserver variability among pathologists, thus leading to over or undertreatment. Computer-aided diagnosis systems can help pathologists reduce overheads and erratic reproducibility. We propose a multi-scale model that detects invasive cancerous areas patterns across the whole slide image. The model extracts tiles of different tissue types at multiple magnification levels and processes them to predict invasive patterns based on local and regional information for accurate T1 staging. Our proposed method yields an F1 score of 71.9, in controlled settings 74.9, and without infiltration 90.0.

11.1 Introduction

Bladder cancer is one of the most commonly diagnosed cancer worldwide, with over 573,000 new cases and 213,000 deaths estimated in 2020 [205]. Approximately 90% of newly diagnosed cases are urothelial carcinomas that arise from the urothelial lining, from these 75% are non-muscle invasive bladder cancers (NMIBC) and 25% are muscle-invasive bladder cancer (MIBC) [109]. NMIBC can be divided into non-invasive papillary tumours Ta, T1 with invasion into the lamina propria and carcinoma in situ Tis, which is a superficial lesion. The three aforementioned groups account for 70, 20, and 10% of NMIBC, respectively [214]. Ta tumours typically are non-aggressive low-risk lesions that can be treated cautiously. However, T1 tumours are considered high-risk lesions with a higher chance of progressing into muscle-invasive tumours, resulting in higher mortality rates [215]. NMIBC accurate staging of the tumour is important to categorize patients into risk groups, and guides which treatment strategy the patient will receive, and, thus, patient outcome [216].

Invasive patterns can be problematic to identify correctly, see Fig 11.1. Morphological features indicating lamina propria (LP) invasion include irregular nests, refraction artifacts, stromal reaction, among others [217]. External factors such as fragmented material and tangential sectioning are often present, providing a false sense of invasion. Also, benign structures like von Brunn nests might mimic invasion [218]. Complex evaluation of multi-variate scenarios leads to high inter-and intraobserver variability, resulting in down- or upstaging [176]. Variability in diagnosing T1 tumours has significant clinical consequences; thus, the necessity for higher reproducibility in assessing staging categories arises.

Computer-aided diagnosis (CAD) systems relying on machine learning methods for medical imaging analysis have proven to be an efficient manner to reduce subjectivity and accelerate the diagnostic procedure [178, 179]. Digital microscopy scanners generate high-resolution digital images from the scanned tissue sections, also named Whole Slide Images (WSI), in a fully automatic way. WSI are stored at multiple magnifications views, allowing the observer to adjust the zoom level, replicating physical microscopes. Pathologists use lower magnification levels to overview the regions of interest and analyse tissue-level morphology, while higher magnification is desirable for observing cell-level morphology.

Convolutional neural networks (CNN) are the gold standard for feature extraction from histological images [118]. CNNs have been used for appli-

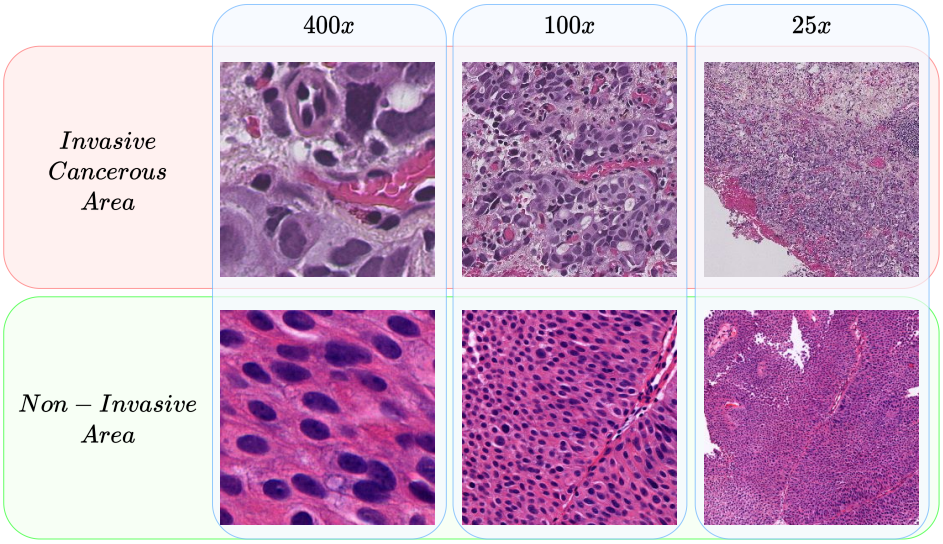


Figure 11.1: Tiles of invasive (T1) and non-invasive (Ta) areas at different magnification levels. The lining contour of the basement membrane of the urothelial layer is interfered with the infiltrating tumour cells. Atypical clusters of tumour cells invade the lamina propria, resulting in irregularly shaped nests.

cations related to pattern detection of cellular features in breast, prostate, among other cancer types [219–222]. There are some examples in the literature where features are extracted at multiple magnification levels to incorporate contextual information of surrounding areas together with detailed information at the cellular level. Both Harmon et al. [76] and Li et al. [75] define tiles at different magnification levels independently from one another. However, Wetteland et al. [223] define one magnification to extract the tiles from and projects the centre of the pixel to other magnifications.

Bladder cancer staging should first and foremost distinguish NMIBC from MIBC. Image modalities such as CT, MRI or PET have proven to be effective, non-invasive methods to identify muscle invasion [224–226]. However, a detailed examination of NMIBC substaging requires histopathological analysis. Automatic applications that help to identify a possible invasion into the lamina propria include tissue segmentation to delineate the boundaries of urothelium, lamina propria and muscle tissue [223, 227]. Yin et al. [34] propose a method to classify tumour regions of H&E stained slides as stage Ta or T1 tumours using machine learning. The method was developed and evaluated in a control set containing patches of

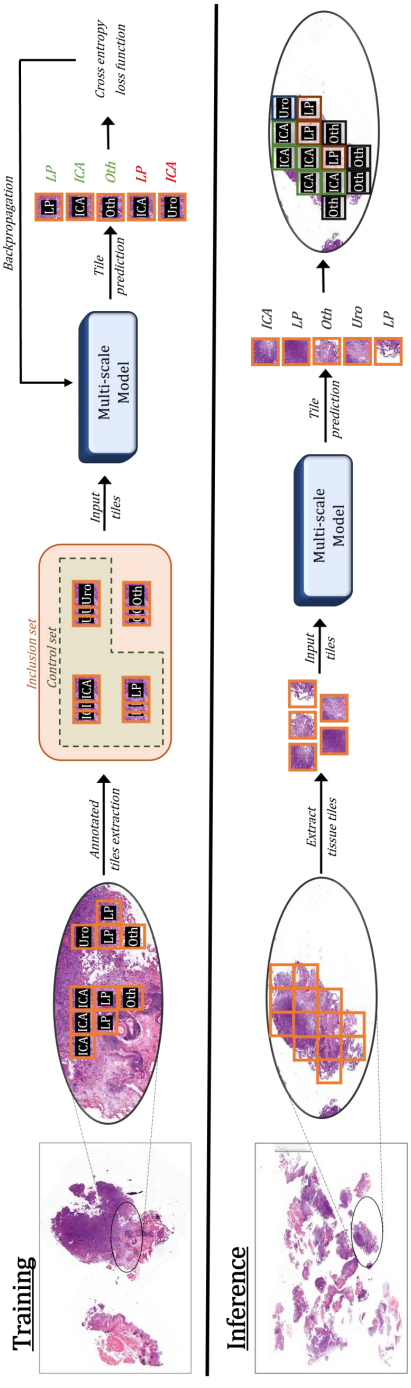


Figure 11.2: During the training stage, exclusively tiles from annotated areas are extracted. Areas from all tissue types are extracted for analysis at the inference stage, resulting in a labelled colormap based on tile predictions.

already segmented and verified Ta and T1 tumour tissue, disregarding the remaining tissue in WSIs. Analysing the entirety of the tissue in the slides is important to define the extent of invasion. Identifying focal points of invasion would provide important information to clinicians as an outcome predictor [228].

We propose an algorithm that finds invasive cancerous areas (ICA) associated with T1 stage across the WSI without the need of previous region of interest (ROI) or tumour segmentation. In ICA, the tumour has spread into the lamina propria that separates the urothelium layer from the bladder muscle beneath, whereas non-invasive areas are the regions of urothelium and lamina propria where no invasion is present. The proposed algorithm takes tissue tiles from all tissue types present in the WSI and discerns invasive from non-invasive patterns using both local and regional information extracted at different magnification levels.

11.2 Methods

11.2.1 Datasets

51 WSI of NMIBC patients were collected from two independent cohorts from Erasmus MC (EMC), Rotterdam, The Netherlands and Stavanger University Hospital (SUH), Stavanger, Norway. The EMC dataset is a multi-center cohort containing 37 H&E stained WSI of high-risk NMIBC. WSI were scanned using a 3DHistech P1000 scanner at 800x magnification stored as MRXS files. The SUH dataset contains 14 WSI of NMIBC that were digitized using a Leica SCN400 scanner at 400x magnification stored in the SCN file format, 8 H&E and 6 HES stained. The combined dataset included 23 non-invasive and 28 invasive tumours. All slides were partially annotated and revised by pathologists, with the annotations serving as ground truth. Only representative regions were annotated from each WSI; thus, no slide was fully annotated. Annotated regions from EMC included mainly tissue types (urothelium, lamina propria, blood, muscle and artifacts), while SUH annotations contain primarily urothelium grading. In some cases, these tissue annotations might come with a subclass, such as grading of urothelium, presence of tumour infiltrating lymphocytes (TILs) in both urothelium and lamina propria, ICA, artifact type, among others. A dataset consisting of tiles was extracted from the annotated areas. Tiles were categorized into four main classes: invasive cancerous areas (*ICA*);

"non-invasive urothelium" (*Uro*); *"lamina propria with no invasion"* (*LP*); and the rest of the tissue types, which we will refer as *Others*. We define non-invasive areas as the regions of *Uro* and *LP* where no invasion is present. *Others* includes blood, muscle and damaged tissue. Further details on the number of tiles extracted from our datasets can be found in Table 11.1.

Table 11.1: Number of WSI and tiles from EMC and SUH datasets.

	WSI	**Tiles**			
		ICA	Urothelium	Lamina Propria	Others
	Ta / T1		with / without TILs	with / without TILs	
EMC	16 / 21	14253	603 / 8859	2816 / 22154	34666
SUH	6 / 8	1333	0 / 60378	1858 / 1054	2726

From the extracted tiles, we defined two subsets; i) a *control set*, and ii) an *inclusion set*. A control set is sampled from the original dataset, containing ICA and selected non-invasive areas (*Uro*, *LP*). Extracting carefully selected tiles, we remove the surrounding tissue present in the slides while maintaining tiles related to the ROI. An inclusion set is defined including *Others* tissue types in addition to ICA, *Uro* and *LP* from the control set. The sets were split into training and test patients, ensuring no data leakage.

11.2.2 Multi-scale Model

We propose a CNN-based deep learning algorithm that can process all available tissue areas in a WSI and produces a segmentation map to detect ICA. Fig. 11.2 depicts the pipeline of the proposed method. The algorithm consists of a combination of feature extractors with input tiles at different magnification levels to obtain detailed as well as contextual information. Then, feature maps are downsampled using Global Average Pooling (GAP) to form a low-dimensional embedding. This information is later concatenated to embeddings from other magnification views and fed into the dense layers which will give the final prediction, as shown in Fig. 14.2. Our proposed algorithm is a TRI-scale model which analyses tiles at three magnification levels (400x, 100x, 25x), inspired by [223].

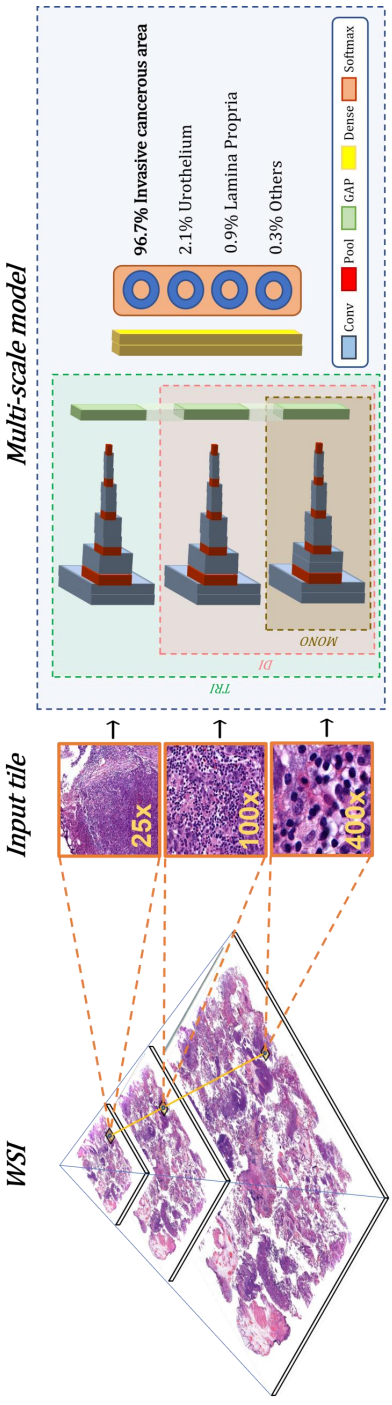


Figure 11.3: Input tiles at different magnification levels are extracted from the WSI and fed into the feature extractors. Extracted features are concatenated and fed into the dense layers which will give the final prediction.

11.3 Experiments

Multiple models using different combinations of magnification levels were trained and tested. The following combinations of magnification views were used: MONO (400x), DI (400x, 100x), and TRI-scale (400x, 100x, 25x). Tiles of size 256×256 were extracted from annotated regions at 400x magnification. Tiles at different magnifications were extracted so that the center of the tile remains the same for all views, as described in [120]. An example of this can be seen in Fig. 11.1. Extracted tiles at 400x are required to be covered by at least 70% of a tissue mask in order to be assigned such label. Background tiles were excluded using a thresholding technique. The number of annotated regions of ICA were significantly less than others. To address the data imbalance problem, we used undersampling of overrepresented classes. Tiles were converted to grayscale as datasets come from different laboratories, with different scanners and stain compositions, and the dataset was not large enough to support color variation. VGG16 was used as a feature extractor. To avoid overfitting, frozen weights were used due to the limited amount of training samples. Early stopping and learning rate decay were enabled. Cross entropy was defined as loss function. Models were trained for a maximum of 50 epochs. All models are implemented in Python 3.6 using Tensorflow machine learning library [229].

The experiments carried out are described as follows: i) for the first set of experiments, denoted E_{mag}, MONO, DI and TRI-scale models are evaluated to determine the relevance of contextual tissue morphology for discerning invasion; ii) for the second set of experiments, denoted E_{ds}, one of the original subsets was used, either control or inclusion, to ascertain the performance variation for adding non-ROI tissue; iii) for the third set of experiments, denoted E_{tils}, we excluded all tiles which included *"tumour infiltrating lymphocytes"* (TILs). TILs are small immune cells that can be found either in the urothelium or the lamina propria. TILs may lead to confusion for discerning them from invasion; iv) finally, for the fourth set of experiments, denoted E_{emc}, we used tiles from one of the hospitals exclusively. We chose EMC over SUH since the number of available data from EMC is far greater, especially regarding the number of ICA tiles. E_{emc} will show how impactful data balance among classes and domain shift are toward detecting invasive cancerous patterns. Also, the experiments were run in a binary and multi-class manner to assess whether grouping of non-invasive tissue improves ICA detection performance.

Table 11.2: Results of the control set are depicted with precision, recall and F1 score for the ICA class only. The experiment names explained in the text are indicated as the method (E_{mag}) and datasets (E_{ds}, E_{tils}, E_{emc}).

	Control set ($\mathbf{E_{ds}}$)					
	ICA vs (Uro+LP+Others)			**ICA vs Uro vs LP vs Others**		
Method ($\mathbf{E_{mag}}$)	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
MONO (400x)	48.5	78.4	60.0	53.8	70.1	60.8
* E_{tils}	69.3	81.2	74.8	67.6	79.9	73.2
* E_{emc}	30.4	90.4	45.5	33.9	83.6	48.3
DI (400x,100x)	65.4	76.7	70.6	64.3	81.1	71.7
* E_{tils}	91.7	79.1	84.9	87.6	87.2	**87.4**
* E_{emc}	41.5	91.3	57.0	42.7	98.2	59.6
TRI (400x,100x,25x)	66.4	81.3	73.1	70.5	79.8	74.9
* E_{tils}	93.8	86.4	**90.0**	85.5	89.1	87.3
* E_{emc}	48.1	94.1	63.6	49.2	93.3	64.4

Table 11.3: Results of the inclusion set are depicted with precision, recall and F1 score for the ICA class only. The experiment names explained in the text are indicated as the method (E_{mag}) and datasets (E_{ds}, E_{tils}, E_{emc}).

	Inclusion set ($\mathbf{E_{ds}}$)					
	ICA vs (Uro+LP+Others)			**ICA vs Uro vs LP vs Others**		
Method ($\mathbf{E_{mag}}$)	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
MONO (400x)	41.0	75.6	53.2	44.3	77.5	55.3
* E_{tils}	69.3	74.3	67.0	52.9	80.3	63.8
* E_{emc}	24.9	91.1	39.1	30.5	80.8	44.3
DI (400x,100x)	64.6	74.3	69.1	59.6	79.1	68.0
* E_{tils}	91.7	78.2	81.9	75.3	86.0	80.3
* E_{emc}	32.3	90.6	47.6	40.3	97.5	57.0
TRI (400x,100x,25x)	65.9	79.2	71.9	63.4	80.3	70.8
* E_{tils}	93.8	81.2	**85.0**	77.6	88.9	**82.9**
* E_{emc}	53.1	79.5	63.7	43.1	93.2	58.9

11.4 Results & Discussion

Results of all experiments are collected in Tables 11.2 and 11.3 for space efficiency. For experiment E_{mag}, TRI-scale models provide the best results for ICA detection, inferring that regional context derived from tissue morphology is beneficial in the task of discerning invasive patterns over non-invasive ones. Hence, models that incorporate lower magnification views provide better performance than those who focus on local patterns solely.

With respect to experiment E_{ds}, results on the control set demonstrate that discerning invasive from non-invasive patterns is possible, carefully selecting ROIs containing urothelium and lamina propria, with and without invasion. We observe that trained models with all sorts of tissue provide comparable results, using the inclusion set. It is, however, with models trained on the inclusion set that we can directly deploy the algorithm for WSI analysis. Models trained with the control set require a pre-processing step to sample carefully selected patches from the slides. Moreover, we further evaluate the discriminatory competence of our models for both sets representing a scatterplot of a two-component feature embedding visualization using t-SNE [230], see Fig. 11.4. Sample embeddings are clustered into separate locations of the feature space based on the class label; hence there is a clear separation of the classes. With regard to the number of classes, binary classification has proven to be more efficient than multi-class for ICA predictions in the inclusion set, but not for the control set. The best performing binary model of the inclusion set, TRI_{incl} binary model, was used to produce probability heatmaps over WSIs from the test set, as shown in Fig. 11.5. Tile predictions were overlaid on the WSI, ranging from blue to red for ICA likelihood. A pathologist could later use the generated heatmaps as a tool for guidance in detecting potential ICA.

Regarding experiment E_{tils}, we noticed that the presence of TILs dampers the predictive performance, as models struggle to differentiate such areas from ICA. Tiles where TILs are present are often missclassified for ICA. Excluding areas with abundant TILs results in a significant F1 score improvement. The positive effect of TILs exclusion is relevant for both sets and all combinations of magnification views, although it is higher when using TRI-scale models. For the TRI_{incl} binary model, the F1 score difference reaches over 13%. Based on the results obtained, a future pipeline should include a posterior classifier to discriminate ICA false positives with infiltration from genuine ICA.

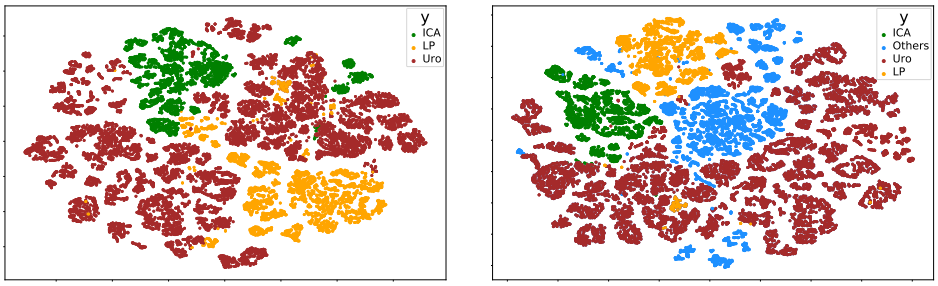


Figure 11.4: tSNE scatterplot of test samples embeddings on the control and inclusion sets, respectively. Embeddings were extracted as the output of the feature extractor and reduced to two components.

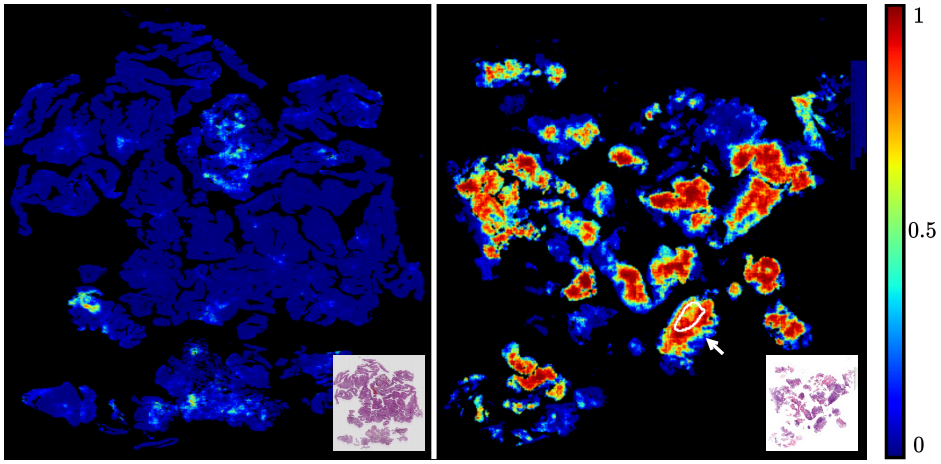


Figure 11.5: Heatmaps showing the probability of a patch to belong to the ICA class, based on the TRI_{incl} binary model. Two test WSI samples with stage Ta (left) and T1 (right) are shown. An annotated ICA region is highlighted by a white border in the T1 WSI. Inclusion set models support the analysis of the entire WSI, not limited to pre-defined ROIs. A thumbnail of the original WSI is displayed at the bottom right corner.

Finally, for experiment E_{emc}, despite reducing scanning and staining variability, we lit upon poor performance in comparison to the other experiments. This can be due to the lack of a substantial number of urothelium tiles, which are present mostly in the excluded cohort, see Table 11.1. Even though most of the ICA tiles remained in the dataset, discarding the majority of the *Uro* class proved unfavourable. Precision results infer a bias towards ICA. Distinguishing ICA patterns for non-invasive patterns

relies mainly on telling Ta from T1 regions apart. As the tumour grows from the urothelial layer into the lamina propria, it is essential to have a significant representation of *Uro* over *LP*.

11.5 Conclusion & Future Work

A multi-scale model was developed for detecting invasive patterns in ICA across the WSI. A control set shows that CNNs distinguish invasive areas from other tumour regions, while including non-diagnostically relevant regions barely decreases the model capability of discerning invasion. Multi-scale models give best performance indicating that using contextual information combined with local patterns is highly beneficial. Predicted heatmaps for the invaded areas can be represented for user-friendly interpretation. Automatic ICA detection can also can be a first step for automated staging of T1 tumours. This can be useful for treatment planning, both alone and in combination with automated grading. The results are promising but due to the lack of large scale manual annotated regions, it is conducted over a small dataset. Future developments should be conducted over larger cohorts and with implemented color normalization schemes to alleviate scanning and staining variability. Likewise, a post-processing step for TILs exclusion should be adopted to alleviate infiltration misclassifications.

Paper 4: NMGrad: Advancing Histopathological Bladder Cancer Grading with Weakly Supervised Deep Learning

NMGrad: Advancing Histopathological Bladder Cancer Grading with Weakly Supervised Deep Learning

S. Fuster[1], U. Kiraz[2,3], T. Eftestøl[1], E. A. M. Janssen[2,3], K. Engan[1]

[1] Dept. of Electrical Engineering and Computer Science, University of Stavanger, 4021 Stavanger, Norway

[2] Dept. of Pathology, Stavanger University Hospital, 4011 Stavanger, Norway

[3] Dept. of Chemistry, Bioscience and Environmental Engineering, University of Stavanger, 4021 Stavanger, Norway

Under review. Preprint available on arXiv:

https://arxiv.org/abs/2405.15275

Abstract:
The most prevalent form of bladder cancer is urothelial carcinoma, characterized by a high recurrence rate and substantial lifetime treatment costs for patients. Grading is a prime factor for patient risk stratification, although it suffers from inconsistencies and variations among pathologists. Moreover, absence of annotations in medical imaging difficults training deep learning models. To address these challenges, we introduce a pipeline designed for bladder cancer grading using histological slides. First, it extracts urothelium tissue tiles at different magnification levels, employing a convolutional neural network for processing for feature extraction. Then, it engages in the slide-level prediction process. It employs a nested multiple instance learning approach with attention to predict the grade. To distinguish different levels of malignancy within specific regions of the slide, we include the origins of the tiles in our analysis. The attention scores at region level is shown to correlate with verified high-grade regions, giving some explainability to the model. Clinical evaluations demonstrate that our model consistently outperforms previous state-of-the-art methods.

13.1 Introduction

Bladder cancer, a prevalent urological malignancy, poses significant clinical challenges in terms of both diagnosis and prognosis [231]. Non-muscle-invasive bladder cancer (NMIBC) accounts for approximately 75% of the newly diagnosed cases of urothelial carcinoma. NMIBC is particularly known for its variable outcomes, necessitating accurate and consistent grading for optimal patient management [109]. The 2022 edition of the European Association of Urology guidelines on NMIBC recommends a stratification of patients into risk groups based on the risk of progression to muscle-invasive disease [10]. Grade, stage, and various other factors contribute to the risk. Precise risk assessment is vital in the management of NMIBC, since treatment strategies not only rely on the presence of muscle invasion.

Grading is based on assessment of the cellular morphology abnormalities of urothelial tissue. In 2004, the WHO introduced a grading classification system (WHO04) for NMIBC based on histological features. WHO04 encompasses three categories: papillary urothelial neoplasm of low malignant potential (PUNLMP), non-invasive papillary carcinoma low grade (LG), and high grade (HG), ranging from lower to higher malignancy respectively [232]. HG is related to lower differentiation, loss of polarity, and pleomorphic nuclei, among others. The intricate evaluation of heterogeneous scenarios contributes to significant inter- and intraobserver variability. Disparities potentially lead to misclassification, and consequently, to inappropriate treatment decisions [176]. WHO04 grading system was subsequently retained in the updated 2016/2022 WHO classifications [233]. However, according to several multi-institutional analyses of individual patient data, the proportion of tumors classified as PUNLMP (WHO 2004/2016) has markedly decreased to very low levels in the last decade. This trend resulted in suggestions to reassess PUNLMP tumors as LG [234, 235]. Therefore, the upcoming grading system will reasonably undergo a modification, shifting towards the inclusion of only LG and HG categories. In recent years, the integration of deep learning techniques within the field of computational pathology (CPATH) has offered promising avenues for enhancing the precision of computer-aided diagnosis (CAD) systems and elucidating discrepancies among pathologists [123, 207]. Consequently, the synergy between histological expertise and CAD technologies is vital for accurate grading assessments.

CPATH is the field of pathology that leverages the potential of CAD

systems to thoroughly analyze high-resolution digital images known as Whole Slide Images (WSIs) for diverse diagnostic and prognostic purposes [118]. WSIs are produced by slide scanners and pre-stored at different magnification levels, emulating the functionality of physical microscopes. Lower magnification is suited for tissue-level morphology examination, while higher magnification for cell-level scrutiny [236]. WSIs are characterized by their substantial size, which can introduce adversarial noise. Bladder cancer WSIs present unique challenges due to their disorganized nature and the presence of diagnostically non-relevant tissue. These slides often include artifacts, such as cauterized or stretched tissue [126]. Moreover, tissue such as blood, muscle and stroma are less informative for grading a tumor. Therefore, the absence of annotations presents a significant challenge for identifying regions of interest (ROIs) [123]. It is crucial to distinguish between region-based labels (e.g., tissue type, grade) and WSI-based labels, including follow-up information and overall patient grade. Grade exemplifies a label that encompasses both perspectives [237, 238]. While clinical reports assign the worst grade observed to a patient, a WSI may exhibit diverse urothelium regions with normal, LG, and HG. This dual nature underscores the complexity of label interpretation when considering both the medical, WSI-focused perspective and the more technical perspective involving regional data analysis and processing.

CPATH is amid a transformative era, aiming to reshape the landscape of digital pathology as we know it [239]. Among the diverse practices in the field, imaging methodologies rooted in convolutional neural networks (CNNs) have emerged as the foundation of feature extraction from histological images [118]. These deep learning networks possess a remarkable capacity to automatically discern morphological and cellular patterns within WSIs [45, 240–244]. Ultimately, CNNs contribute to more precise and timely clinical decisions. However, training deep learning models in CPATH presents challenges when only WSI-level labels are available, lacking region-based annotations [123, 245–247]. To address these, weakly supervised learning techniques, like attention-based methods and multiple instance learning (MIL), are employed. However, MIL methods can be susceptible to individual instances dominating the weighted aggregation of the WSI representation [184, 194, 248]. In the context of WSIs, the tissue is distributed across the slide, for which regions typically present similar features within and pathologists pinpoint ROIs with crucial information [249]. Specifically while grading, situations may arise where multiple instances in close proximity exhibit HG characteristics, while other regions

may concurrently display LG attributes. As a result, constraining instances to specific regions enhances our understanding of the diverse features within WSIs. Consequently, a conventional MIL architecture approach might not be appropriate, because there is a susceptibility to information leakage between regions. A model accommodating for the nested structure of WSIs, wherein tissues are part of a region, and regions, in turn, belong to a WSI, may more effectively capture the clinical WSI-level grading label [173, 174, 250].

In this study, we introduce a novel pipeline for grading NMIBC using histological slides, referred to as nested multiple grading (NMGrad). The proposed solution starts by tissue segmentation of the WSI, separating urothelium from other tissue types. The next step categorizes extracted tiles of urothelium areas into location-dependent regions for predicting the patient's WHO04 grade. We implemented a weakly supervised learning framework using attention mechanisms, and a nested aggregation architecture for ROI differentiation. Our method offers an innovative approach for generating diagnostic suggestions, with generated heatmaps for highlighting tiles and ROIs independently.

13.2 Related Work

Numerous studies in the domain of computational pathology for bladder cancer diagnostics have emerged in recent years [23]. In Wetteland et al. [31], a pipeline for grading NMIBC is introduced. This pipeline identifies relevant areas in the WSIs and predicts cancer grade by considering individual tile predictions and applying a decision threshold to determine the overall patient prediction. Their results demonstrate promising performance, with potential benefits for patient care. In Zheng et al. [181], the authors focus on the development of deep learning-based models for bladder cancer diagnosis and predicting overall survival in muscle-invasive bladder cancer patients. They introduce two deep learning models for diagnosis and prognosis respectively. They show that their presented algorithm outperformed junior pathologists. In Jansen et al. [175], the authors propose a fully automated detection and grading network based on deep learning to enhance NMIBC grading reproducibility. The study employs a U-Net-based segmentation network to automatically detect urothelium, followed by a VGG16 CNN network for classification. Their findings demonstrate that the automated classification achieves moderate agreement with consensus and individual

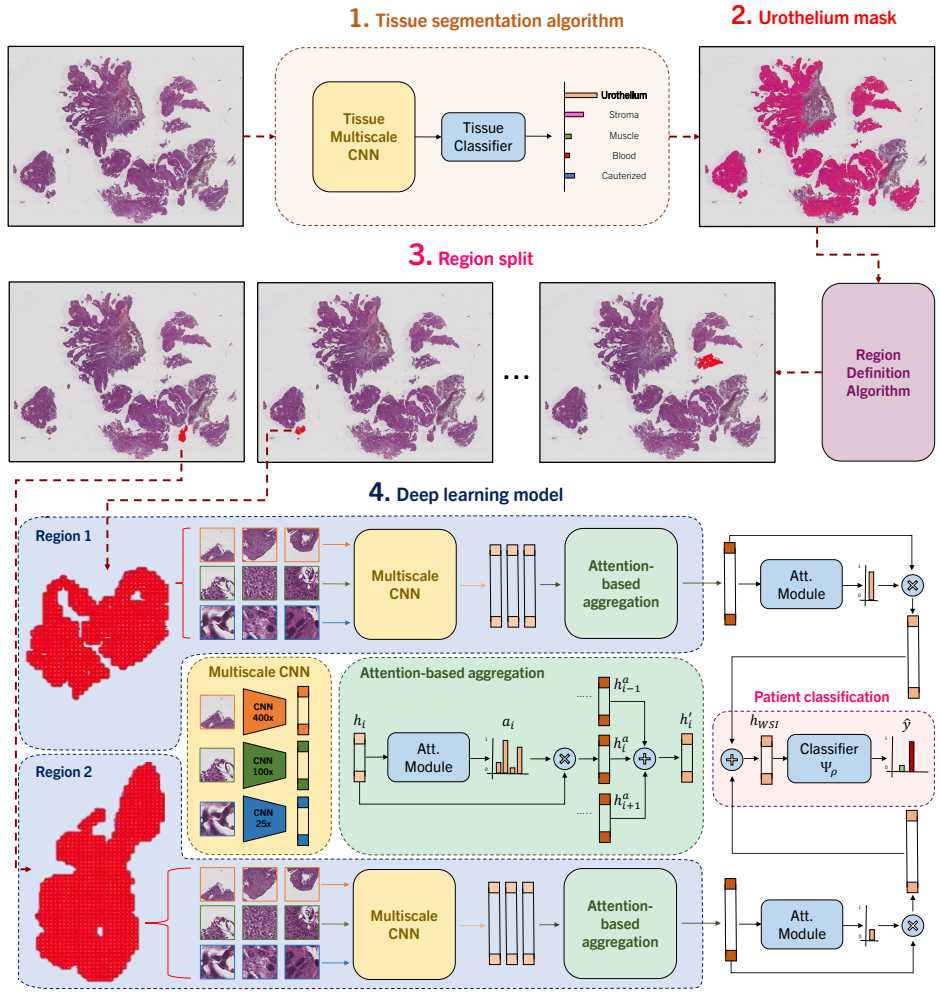


Figure 13.1: NMGrad pipeline. Initially, we apply a tissue segmentation algorithm for ROI extraction Then, we pinpoint diagnostically significant urothelium areas within WSIs. Subsequently, we split the urothelium mask into regions, based on proximity and size, and extract tile triplets. In a hierarchical fashion, we further transform these triplets within their corresponding regions into region feature embeddings using an attention-based aggregation method. All the region representations are then consolidated into a comprehensive WSI-level representation through a weight-independent attention module. Finally, this WSI feature embedding is input into the WHO04 grading classifier in order to produce accurate WSI grade predictions.

grading from a group of three senior uropathologists. Spyridonos et al. [251]

investigate the effectiveness of support vector machines and probabilistic neural networks for urinary bladder tumor grading. The results indicate that both SVM and PNN models achieve a relatively high overall accuracy, with nuclear size and chromatin cluster patterns playing key roles in optimizing classification performance.

Zhang et al. [252] address a common limitation of interpretability in CAD methods. To tackle this, they introduce MDNet, a novel approach that establishes a direct multimodal mapping between medical images and diagnostic reports. This framework consists of an image model and a language model. Through experiments on pathology bladder cancer images and diagnostic reports, MDNet demonstrates superior performance compared to comparative baselines. Zhang et al. [33] propose a method that leverages deep learning to automate the diagnostic reasoning process through interpretable predictions. Using a dataset of NMIBC WSIs, the study demonstrates that their method achieves diagnostic performance comparable to that of 17 uropathologists.

Two critical challenges we identified include summarizing information from local image features into a WSI representation, and the scarcity of annotated datasets. Effectively translating detailed local information to the WSI level is complex, particularly in tasks like grading NMIBC. Moreover, the limited availability of well-annotated datasets hinders the development and evaluation of robust models. To tackle these issues, weakly supervised methods have emerged as a standout tool in CPATH [123]. While weakly supervised methods are widespread, some studies still rely on annotations and supervised learning. However, there is a growing consensus for the future of CPATH to predominantly embrace weakly supervised approaches. This shift is driven by the impracticality of obtaining detailed annotations for large datasets covering various cancers and tasks. Among the various weakly supervised methods, attention-based MIL (AbMIL), a popular instance aggregation method, exploits attention mechanisms in order to mitigate the uncertainty from individual instances and enhance interpretability [84, 253]. AbMIL bridges the gap between limited supervision and the spatial details necessary for accurate analysis and explainability. An evolution of MIL model architectures relies on the arrangement of the data within bags, where instances are further subdivided into finer groups. This concept is referred to as nesting [173, 174]. Nested architectures preserve a sense of localization or categorization by selectively processing data instances within individual subgroups. Subsequently, they

aggregate summarized information from the subgroups into a final bag representation.

In our work, we aim to bridge the gap between non-annotated datasets, weakly supervised methods and the intrinsic categorization of WSI data. Therefore, we leverage the nested MIL with attention mechanisms (NMIA) model architecture we proposed in Fuster et. al. [250], for accurate and interpretable NMIBC grading. Finally, in order to overcome the lack of annotations for defining the tissue of interest, a tissue segmentation algorithm TRI-25x-100x-400x was proposed by our research group in [26]. More recent works from our research group on tissue segmentation have found adoption within the scientific literature [27, 254]. The utilization of this segmentation algorithm offers the opportunity to extract tiles specifically from the urothelium, contributing to a refined and targeted extraction process.

13.3 Data Material

The dataset comprises a total of 300 digital whole-slide images (WSIs) derived from 300 patients diagnosed with NMIBC, from the Department of Pathology, Stavanger University Hospital (SUH) [237, 255]. The glass slides were digitized using a Leica SCN400 slide scanner, saved in the vendor-specific SCN file format. Collected over the period spanning 2002 to 2011, this dataset encompasses all risk group cases of non-muscle invasive bladder cancers. The biopsies were processed through formalin-fixation and paraffin-embedding, and subsequently, 4µm thick sections were prepared and stained using Hematoxylin, Eosin, and Saffron (HES). Furthermore, all WSIs underwent meticulous manual quality checks, ensuring the inclusion of only high-quality slides with minimal or no blur. Due to the cauterization process used in the removal of NMIBC, some slides may exhibit areas with burned and damaged tissues. All WSIs originate from the same laboratory, resulting in relatively consistent staining color across the dataset.

All WSIs were graded by an expert uropathologist in accordance with the WHO04 classification system, as either LG or HG, thus providing slide-level diagnostic information. However, the dataset lacks region-based annotations pinpointing the precise areas of LG or HG regions within the WSI. Consequently, the dataset is considered weakly labeled. For WSIs labeled as LG, at least one LG region is expected, with the possibility of presenting non-cancerous tissue in other regions. As for HG slides, at least

one region should display HG tissue, while other regions may exhibit a LG appearance or non-cancerous tissue. Given the absence of alternative gold standards, we are compelled to continue utilizing a grading assessment that may have limitations for training and evaluating our algorithms. The dataset employed in this study was divided into three subsets: 220 WSI/patients for training, 30 for validation, and 50 for test. The split employed ensures that each subset maintains the same proportional representation of diagnostic outcomes. This stratification encompassed factors such as WHO04 grading, cancer stage, recurrence, and disease progression to best mirror the diversity of the original data material. The distribution of LG and HG WSIs within each dataset is detailed in Table 1 for reference.

Within a subset of the test set, denoted as $\text{Test}_{\text{ANNO}} \in \text{Test}$, 14 WSIs contain either one or two annotated regions of confirmed LG or HG tissue, verified by an expert uropathologist. It is noteworthy that not all regions were annotated. The labels of these regions correspond to the associated weak label of the WSI.

	Low-grade	High-grade
Train	124 (0)	96 (0)
Validation	17 (0)	13 (0)
Test	28 (7)	22 (7)

Table 13.1: Number of slides and corresponding grade per subset. The number between parenthesis corresponds to slides containing some annotations giving region-based labels.

13.4 Methods

We propose NMGrad, a pipeline that begins with a tissue segmentation algorithm for extracting urothelium tissue. Subsequently, the urothelium is divided into localized regions. Thereafter we employ a weakly supervised learning method to predict tumor grade from the segmented urothelium regions. We exploit the sense of region locations by adopting a nested architecture with attention, NMIA [250]. The rationale behind employing this structured data arrangement analysis is to identify relevant instances and regions within the WSI. The attention mechanism and the nested bags/regions also contributes to a more precise and insightful analysis of the data. An overview of NMGrad can be visualized in Fig. 1.

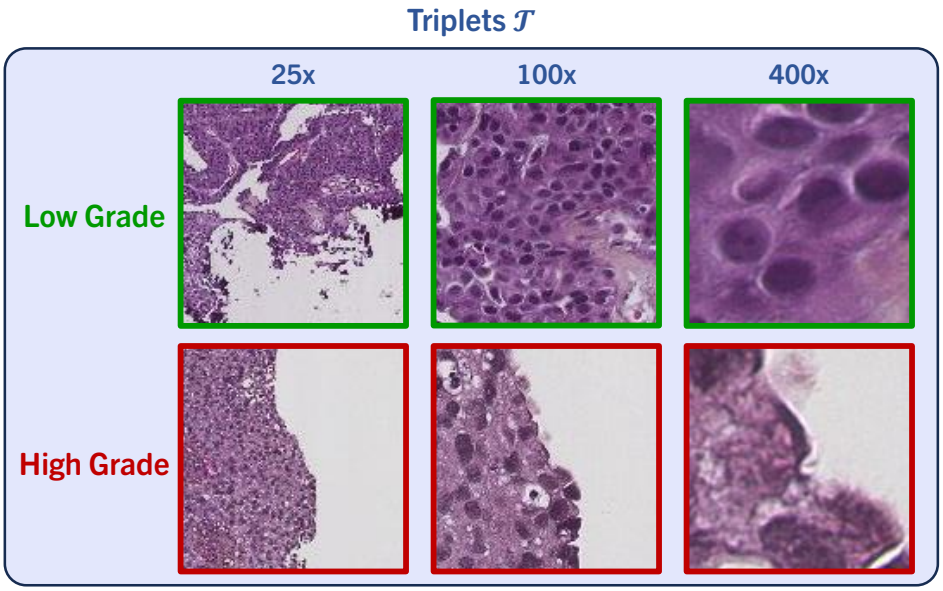


Figure 13.2: We obtain comprising sets of three tiles at different magnification levels named triplets $\mathcal{T}$, enabling detailed examination. Tile triplets demonstrate regions associated with low- and high-grade features.

13.4.1 Automatic Tissue Segmentation & Region Definition

We utilize the tissue segmentation algorithm introduced by Wetteland et al. [26] to automatically generate tissue type masks, facilitating the subsequent extraction of tiles. We define triplets $\mathcal{T}$ of tissue, which consist of a set of three tiles at various magnifications levels, namely 25x, 100x and 400x. An example is shown in Fig. 2. We used a tile size of 128×128 for all magnification levels. Triplets are formed to maintain consistency, ensuring that the center pixel in every tile accurately represents the same physical point. The tissue segmentation algorithm works at tile level, and classifies all triplets $\mathcal{T}$ in the WSI as $y \in \mathcal{Y} = \{$*urothelium, lamina propria, muscle, blood, damage, background*$\}$. As grading relies on *urothelium* alone, we utilize the *urothelium* mask for defining valid areas for tile extraction, as described in [120]. In this work, various magnifications are explored for defining the model's input, either using mono-scale MONO (400x), di-scale DI (400x, 100x) or tri-scale TRI (400x, 100x, 25x). We employed 400x magnification to establish a tight grid of tiles for data extraction

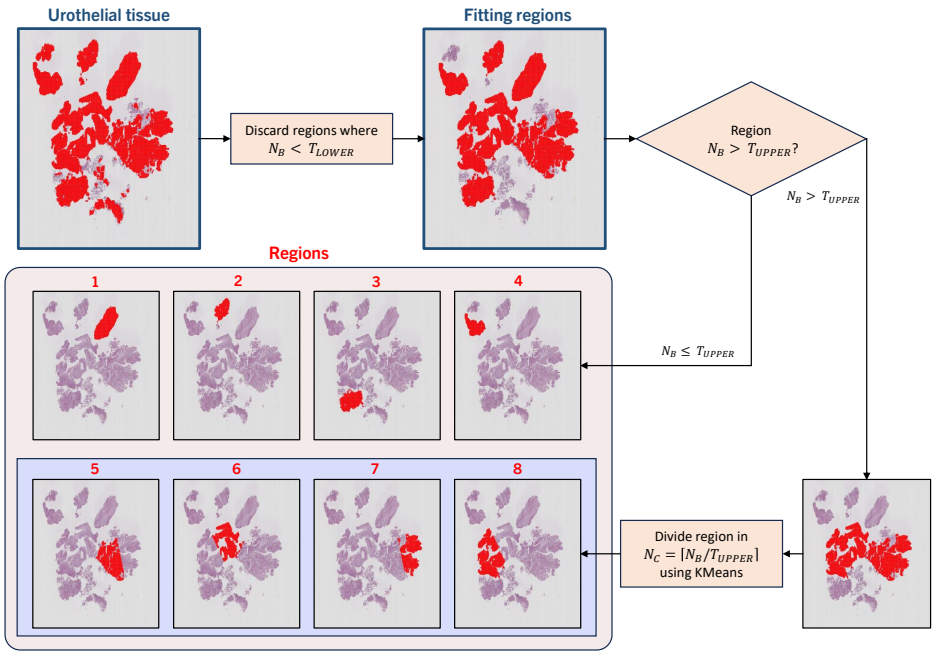


Figure 13.3: Region Definition. Urothelial tissue within a WSI is eligible for tile extraction. Blobs of tiles are formed, and blobs smaller than a threshold T_{LOWER} are discarded. From the remaining blobs, any smaller than T_{UPPER} are kept and defined as a region. For blobs bigger than T_{UPPER}, the blobs is subdivided into smaller pieces using the location of individual tiles within and KMeans clustering. The obtained clusters are designated as regions.

purposes. These sets of tiles are later fed to the grading models, where each magnification tile is processed by its respective weight-independent CNN. To preserve the sense of location within the image, we define regions. This results in the following stratified division of data: all urothelium in the WSI, scattered regions of urothelium, and finally, individual tiles of urothelium.

Region Definition

For defining regions out of the extracted urothelium tiles, we define blobs of tiles $\text{URO}_{\text{BLOB}} \subseteq \text{URO}$. URO_{BLOB} is formed when tiles are 8-connected, and this joint set of tiles is the representation of a region $\text{URO}_{\text{BLOB}} = \{\mathcal{T}_1, \mathcal{T}_2 ....\}$. A region is eligible for inclusion if the number of tiles N_B is

higher than a threshold number T_{LOWER}. Any blob with $N_B < T_{\text{LOWER}}$ tiles is discarded, along with the tiles within. As NMIBC WSI can contain large tissue bundles, resulting in sizable blobs, we also define an upper limit threshold T_{UPPER}. For blobs where $N_B \geq T_{\text{UPPER}}$, we split the region into several sub regions for more detailed analysis. We apply KMeans clustering over the coordinates of the tiles x, y within the blob for location sense, defining the number of clusters as $N_C = \lceil N_B / T_{\text{UPPER}} \rceil$. This results in joint regions within a bundle of tissue of consistent size, as observed in regions 5-8 in Fig. 3.

13.4.2 Multiple Instance Learning in a WSI context

Multiple instance learning (MIL) is a weakly supervised learning method where unlabeled instances are grouped into bags with known labels [184, 248]. A dataset $\mathcal{X}, \mathcal{Y} = \{(\mathbf{X}^i, y^i), \forall i = 1, ..., N\}$ is formed of pairs of sample sets $\mathbf{X}$ and their corresponding labels y, where i denotes a bag index. In the context of WSI, the bag can be one patient or one WSI or one region. In a conventional MIL data arrangement, we consider the bag $\mathbf{X}$ to be one WSI consisting of instances $\mathbf{x}_l$:

$$\mathbf{X} = \{\mathbf{x}_l, \forall l = 1, ..., L\} \tag{13.1}$$

where L is the number of instances in the bag. In our study, $\mathbf{x}$ refers to individual tiles in a set of extracted tiles from a patient slide $\mathbf{X}$. A feature extractor $G_\theta : \mathcal{X} \rightarrow \mathcal{H}$ transforms image tiles, $\mathbf{x}_l$, into low-dimensional feature embeddings, $\mathbf{h}_l$. At this point, the bag structure previously formed remains intact, as instances have been simply transformed. Given a label y for a WSI, the training objective of the model is to predict the grade observed in the WSI. However, to deduce the specific region(s) within a WSI that leads to the patient's diagnosis of either LG or HG is of utmost importance. This entails the model's capability to discern and highlight the critical areas within the WSI that play a pivotal role in the diagnosis. In order to accomplish this goal, we adopted attention-based multiple instance learning (AbMIL) as our MIL framework, using attention-based aggregation as shown in Fig. 1. An attention score a_i for a feature embedding $\mathbf{h}_i$ can be calculated as:

$$a_i = \frac{\exp\{\mathbf{w}^\top(\tanh(\mathbf{V}\mathbf{h_i}^\top) \odot \text{sigm}(\mathbf{U}\mathbf{h_i}^\top))\}}{\sum_{l=1}^{L} \exp\{\mathbf{w}^\top(\tanh(\mathbf{V}\mathbf{h_l}^\top) \odot \text{sigm}(\mathbf{U}\mathbf{h_l}^\top))\}} \tag{13.2}$$

where $\mathbf{w} \in \mathbb{R}^{L\times 1}$, $\mathbf{V} \in \mathbb{R}^{L\times M}$ and $\mathbf{U} \in \mathbb{R}^{L\times M}$ are trainable parameters and $\odot$ is an element-wise multiplication. Furthermore, the hyperbolic tangent $\tanh(\cdot)$ and sigmoid $\text{sigm}(\cdot)$ are included to introduce non-linearity for learning complex applications. The benefit of attention modules extends beyond interpretability for understanding the model's decision-making process, as it also grants enhanced predictive capabilities by prioritizing salient features. This is because attention scores directly influence the forward propagation of the model, allowing it to focus on the most relevant and informative regions within the data. Once the attention scores $\mathbf{A}$ are obtained, we obtain the patient prediction $\hat{y}$ using a patient classifier Ψ_ρ as:

$$\hat{y} = \Psi_\rho(\mathbf{H}^a) = \Psi_\rho(\mathbf{A} \cdot \mathbf{H}) = \Psi_\rho(\mathbf{A} \cdot G_\theta(\mathbf{X})) \tag{13.3}$$

Nested Multiple Instance Architecture

An evolution of the conventional AbMIL architecture defines levels of bags within bags where only the innermost bags contain instances. This is referred to as nested multiple instance with attention (NMIA) [250]. A bag-of-bags for a WSI $\mathbf{H}_{\text{WSI}}$ contains a set of inner-bags, or regions, $\mathbf{H}_{\text{REG},k}$:

$$\mathbf{H}_{\text{WSI}} = \{\mathbf{H}_{\text{REG},k}, \forall k = 1, ..., K\} \tag{13.4}$$

where the number of inner-bags K varies between different WSI. Ultimately, $\mathbf{H}_{\text{REG},k}$ contains instance-level representations $\mathbf{h}_{\text{TILE},l}$ of tiles located within the physical region. This serves to further stratify into clusters or regions, and accurately represent the arrangement of the scattered data, where tiles belong to particular tissue areas and these themselves to the WSI.

$$\mathbf{h}_{\text{REG}} = \mathbf{A}_{\text{TILE}} \cdot \mathbf{H}_{\text{TILE}} = \mathbf{A}_{\text{TILE}} \cdot G_\theta(\mathbf{X}) \tag{13.5}$$

Finally, the ultimate WSI representation h_{WSI} is fed to the classifier for obtaining the grade prediction leveraging the region representations $\mathbf{H}_{\text{REG}}$ and the attention scores $\mathbf{A}_{\text{REG}}$ obtained from those same representations:

$$\hat{y} = \Psi_\rho(h_{\text{WSI}}) = \Psi_\rho(\mathbf{A}_{\text{REG}} \cdot \mathbf{H}_{\text{REG}}) \tag{13.6}$$

13.5 Experiments

Within the scope of our experimental investigation, we systematically assessed and compared the impact of diverse magnification levels and combinations of them, namely MONO, DI and TRI-scale models. We further investigate the impact of several weakly supervised aggregation techniques in the performance of our deep learning model. These aggregation techniques vary in their ability to distill valuable diagnostic insights from the data, namely mean, max and attention-based. We put special emphasis in the comparison between a standard AbMIL architecture and the nested model proposed in NMGrad. Finally, we compare our solution to current state-of-the-art methods. The code is available at https://github.com/Biomedical-Data-Analysis-Laboratory/GradeMIL.

We list the details of design choice of the models during training. VGG16 is used as CNN backbone, with ImageNet pretrained weights [256]. In preliminary experiments, we explored various architectures and found that VGG16 exhibited favorable performance across multiple tasks on NMIBC WSIs. Stochastic gradient descent (SGD) was set as optimizer with a learning rate of 1e-4, a batch size of 128, and a total of 200 epochs, with 30 epoch for early stopping based on the AUC score on the validation set. A total number of 5000 tiles per WSIs were preemptively randomly sampled to be further sub-sampled during training. Focal Tversky loss (FTL) is employed [257]. The Tversky index (TI) leverages false predictions, emphasizing on recall in case of large class imbalance tuning parameters α and β. TI is defined as:

$$\mathrm{TI}_c(\hat{y}, y) = \frac{1+\sum_{i=1}^{N} \hat{y}_{i,c} y_{i,c} + \epsilon}{1+\sum_{i=1}^{N} \hat{y}_{i,c} y_{i,c} + \alpha \sum_{i=1}^{N} \hat{y}_{i,\hat{c}} y_{i,c} + \beta \sum_{i=1}^{N} \hat{y}_{i,c} y_{i,\hat{c}} + \epsilon} \tag{13.7}$$

where $\hat{y}_{i,\hat{c}} = 1 - \hat{y}_{i,c}$ and $y_{i,\hat{c}} = 1 - y_{i,c}$ are the probability that sample i is not of class $c \in \mathcal{C}$. ϵ is used for numerical stability, preventing zero division operations. FTL employs another parameter γ for leveraging training examples hardship:

$$\mathrm{FTL}_c(\hat{y}, y) = \sum (1 - \mathrm{TI}_c(\hat{y}, y))^{1/\gamma} \tag{13.8}$$

Model	Accuracy	Precision	Recall	F1 Score	κ	AUC
AbMIL$_{\text{MONO}}$	0.68(0.07)	0.71(0.07)	0.68(0.06)	0.67(0.07)	0.36(0.12)	0.81(0.07)
AbMIL$_{\text{DI}}$	0.79(0.09)	0.80(0.10)	0.78(0.09)	0.78(0.09)	0.57(0.18)	0.85(0.13)
AbMIL$_{\text{TRI}}$	0.82(0.07)	0.82(0.07)	0.82(0.07)	0.82(0.07)	0.64(0.14)	0.91(0.04)
MEAN$_{\text{TRI}}$	0.81(0.03)	0.83(0.03)	0.80(0.03)	0.80(0.03)	0.61(0.05)	0.92(0.03)
MAX$_{\text{TRI}}$	0.80(0.06)	0.80(0.06)	0.78(0.06)	0.79(0.07)	0.58(0.13)	0.85(0.06)
NMGrad$_{\text{MONO}}$	0.68(0.09)	0.71(0.08)	0.69(0.08)	0.68(0.09)	0.37(0.16)	0.80(0.06)
NMGrad$_{\text{DI}}$	0.83(0.03)	0.85(0.03)	0.82(0.03)	0.82(0.03)	0.65(0.06)	0.91(0.04)
NMGrad$_{\text{TRI}}$	**0.86(0.03)**	**0.87(0.02)**	**0.85(0.04)**	**0.85(0.03)**	**0.71(0.06)**	**0.94(0.01)**
Wetteland et al. [31]	0.90(-)	0.87(-)	0.80(-)	0.83(-)	-	-
Jansen et al. [175]	0.74(-)	-	0.71(-)	-	0.48(0.14)	-
Zhang et al. [33]	0.95(-)	-	-	-	-	0.95(-)

Table 13.2: Test performance for various aggregation techniques in weakly supervised learning. We provide the average of five runs, with the standard deviation shown in parentheses. The table presents the different approaches employed in the field, including aggregation techniques that involve considering spatial separation of the instances. We also explore the use of multiple magnification levels, considering 400x as the foundation for all magnification analysis. We also shown the results from other bladder cancer grading works, although in other datasets.

13.6 Results & Discussion

The utilization of our proposed pipeline NMGrad using TRI-scale input emerged as a standout performer, as of Table 13.2. TRI-scale models showcased the ability to capture relevant patterns across different magnifications, substantially enhancing the overall grading accuracy. Comparing the effects of scales on the model performance shows a larger gap in performance between MONO and TRI models than structuring the processing of data using NMGrad or a standard AbMIL architecture. Furthermore, our exploration of aggregation techniques extended to mean and max aggregation methods, which do not include in-built attention mechanisms, yielded less promising outcomes. The absence of attention mechanisms rendered these techniques less effective in capturing nuanced features, underscoring the significance of attention mechanisms.

NMIA architecture embedded in NMGrad, which employs attention mechanisms for both tile and region aggregation, marked a substantial leap in performance in comparison to mean and max aggregation. This strategy provided the strengths of attention-based aggregation and ROI localization via a nested architecture. The incorporation of attention mechanisms allowed the model to pinpoint and emphasize critical visual cues within WSIs related to urothelial cell differentiation, ultimately resulting in a notable enhancement in predictive accuracy for bladder cancer grading. Finally, in a direct comparison to the previous best performing model proposed by Wetteland [31], they implemented weakly supervised learning in a naive manner, where all patches were assigned a weak label. Predictions are made at the patch level, and the determination of WSI-level prediction relies somewhat arbitrarily on the summation of patch predictions, neglecting consideration of localized regions. Using our proposed solution $\text{NMGrad}_{\text{TRI}}$, the solution aligns more closely with clinical expectations, such as the presence of one or more regions indicative of HG if the WSI is classified as HG, rather than scattered patches. Furthermore, $\text{NMGrad}_{\text{TRI}}$ capacity to learn attention scores offers interpretability, as opposed to relying on ad-hoc rules for post-processing patches into a final prediction. Ultimately, we obtained slightly better F1 score, with a trade-off on accuracy. We have also adhered the works of Jansen [175] and Zhang [33] for comparing the performance of state-of-the-art grading algorithms, although the results correspond to their in-house cohorts, different from ours. The development of our deep learning model $\text{NMGrad}_{\text{TRI}}$ for predicting the grading of bladder cancer represents a significant advancement in the realm of accurate grading

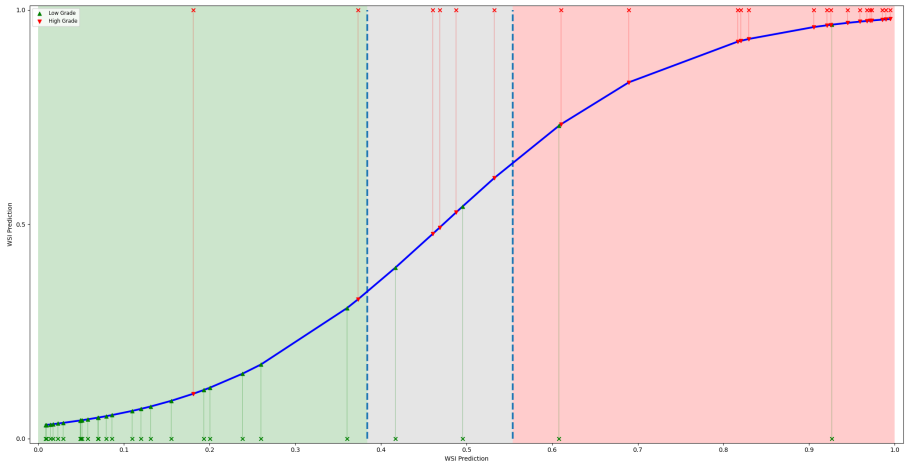


Figure 13.4: Plot displaying the WSI predictions of the test set, with green shading representing the LG confidence interval, red for HG, and gray denoting the uncertainty interval. Additionally, a blue line depicts the regression line fitting the predictions.

of bladder tumors.

In order to enhance the fidelity of binary decisions, we opted to introduce an uncertainty spectrum, thereby introducing a third class. Given the output predictions of test set WSIs, we define the uncertainty spectrum as $[\mu_{\hat{y}(y=LG)} + \sigma_{\hat{y}(y=LG)}, \mu_{\hat{y}(y=HG)} - \sigma_{\hat{y}(y=HG)}]$. A plot illustrating the concept and WSI predictions is shown in Fig. 4. It was observed that by excluding predictions falling within the uncertainty spectrum, the overall F1 score increased to 0.89. This underscores the potential utility of considering skepticism regarding non-confident predictions for robust clinical interpretation.

Due to the attention scores generated at the inference stage, we are able

	Accuracy	**Precision**	**Recall**	**F1 Score**
Attention	0.76	0.81	0.69	0.75
Prediction	0.89	0.83	0.91	0.87

Table 13.3: Region-level prediction and attention correspondence on annotated areas, using $\text{NMGrad}_{\text{TRI}}$. We individually compare the degree of consensus of annotations with both the highest attributed attention within the WSI and the output region prediction of the classifier Ψ_ρ.

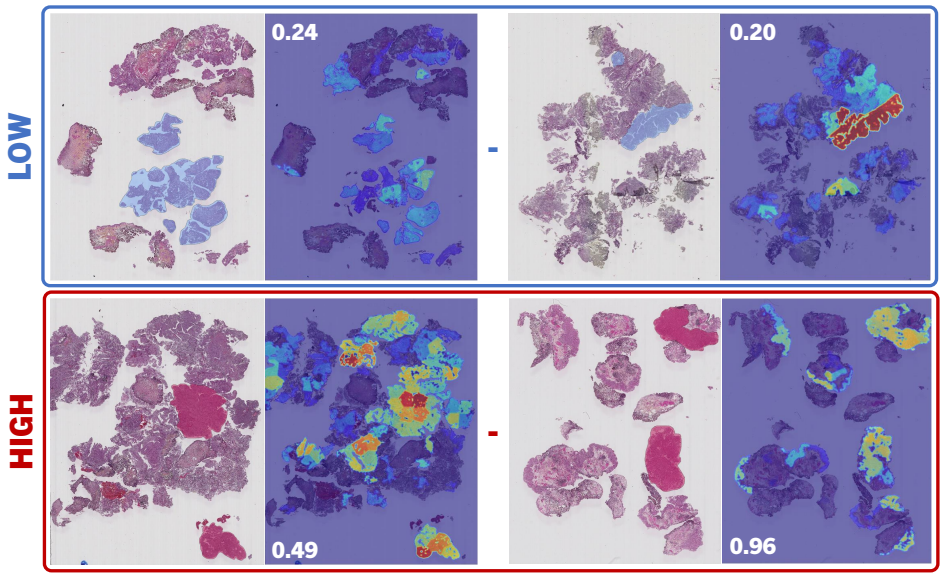


Figure 13.5: Region-level attention score heatmaps. Example regions of annotations of low- and high-grade ROIs annotated by an uropathologist are compared to the output attention provided by the proposed model NMGrad, left to right respectively. The choice of annotated ROIs correspond to highest attention scores, for red and blue correspond to low and high attention correspondingly. We have included the WSI-level prediction score for reference.

to visualize a heatmap, as in Fig. 5. $\text{NMGrad}_{\text{TRI}}$ demonstrated effectiveness in correctly assessing individual tiles and ROIs, despite being trained solely on patient-level weak labels. We consulted the generated heatmaps with experienced pathologists for qualitative analysis. Results on the annotated set of regions, $\text{Test}_{\text{ANNO}}$, exhibited competence to discern LG and HG regions, as of Table 13.3. Region attention scores were considered for direct comparison to region prediction scores utilizing the classifier Ψ_ρ, as values for both scores are restrained between 0 and 1. These two scores were evaluated individually against the annotated areas. It was corroborated that higher attention scores are associated with HG areas in high-grade WSIs. However, the same does not apply for low-grade. For low-grade WSIs, we observed a wide range of possibilities, where high-attention is spread across regions. Ideally, one would anticipate a direct correlation between HG and elevated attention, and conversely, a correlation between

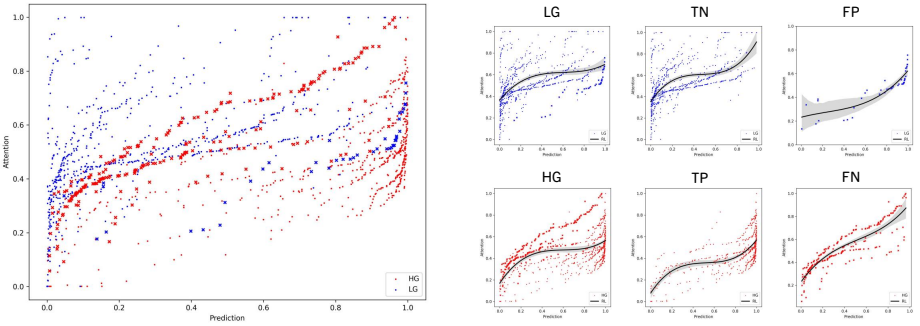


Figure 13.6: Correlation between region attention scores and output prediction of region embeddings on the test set of WSIs. Regions from accurately predicted WSIs (TN, TP) are denoted by squares, while those from incorrectly predicted WSIs (FN, FP) are marked with crosses. A discernible pattern emerges, where low attention scores align with diminished predictions, and conversely, higher attention scores correlate with elevated predictions. Additionally, an observable trend indicates that mispredictions tend to manifest on the opposite end of the spectrum, with low-grade instances concentrating high attention and prediction scores, and vice versa. The trend is represented with a polynomial regression line (RL).

LG and reduced attention. However, this correlation is primarily observable in the positive class (HG), aligning with the inherent design of MIL, which is tailored for identifying positive instances. In contrast to attention, we observed a high degree of correspondence between the label of annotated ROIs and the output region predictions.

We further investigated the correlation between region attention scores and the output region predictions, as displayed in Fig. 6. We observe a generalized reciprocity between low-grade having smaller attention scores and prediction outputs, and vice versa. Moreover, the high-grade range of values is more limited to a lower range compared to the low-grade. For instance, low-grade areas with predictions rounding zero show the broadest range of values. This observation aligns with the earlier statement, wherein the positive class typically exhibits a more focused distribution of attention scores, predominantly linked with positive HG instances. In contrast, the negative class disperses attention across various regions within the WSI despite all presenting similar LG features. In regards to missclassified WSIs, we noted that LG WSIs manifest both high attention and prediction scores, whereas HG slides display a broad range of values. When examining the regression lines of TP and FP, they exhibit similar trends, as do TN and

FN respectively. Essentially, misspredicted WSIs exhibit characteristics contrary to their assigned class.

To augment the evaluation process, we integrate correlation calculations with follow-up information, thereby ensuring a more thorough assessment of our model's performance. We employed Cramér's V correlation coefficient φ_c for calculating the intercorrelation between grading and the event of recurrence and progression, for the test set [258]. We observed a lack of correlation between grading and recurrence for either the manual or automatic grading, aligning with expectations. However, for progression, NMGrad exhibited a higher correlation than the uropathologist (0.32 vs. 0.26, respectively). In accordance with [259], a strong correlation between grading and progression is observed. Notably, these correlations suggest that NMGrad's grade may hold greater predictive value for assessing the likelihood of progression in the context of NMIBC.

13.7 Conclusion

Accurate grading of NMIBC is paramount for patient risk stratification, but it has long suffered from inconsistencies and variations among pathologists. Furthermore, the pathological workload is increasing, as well as its expenses. In response to this challenge, we introduced the NMGrad pipeline, a pioneering approach in bladder cancer grading using WSIs. NMGrad starts by using a tissue segmentation algorithm, finding areas of urothelium in the slides. Thereafter, it leverages a nested AbMIL model architecture to precisely identify diagnostically relevant regions within WSIs and collectively predict tumor grade. Moreover, through a multiscale CNN model, NMGrad processes urothelium tissue tiles at multiple magnification levels. We observed that in high-grade patients, attention scores pinpoint specific ROIs, while in low-grade patients, attention is more dispersed, deviating from the expected MIL pattern. Our clinical evaluations demonstrate that NMGrad consistently outperforms previous state-of-the-art methods, achieving 0.94 AUC score. This achievement represents a significant advancement in the field of bladder cancer diagnosis, with the potential to improve patient outcomes, reduce economic burdens, and enhance the quality of care in the management of this challenging disease.

Paper 5: Self-Contrastive Weakly Supervised Learning Framework for Prognostic Prediction Using Whole Slide Images

Self-Contrastive Weakly Supervised Learning Framework for Prognostic Prediction Using Whole Slide Images

S. Fuster[1], F. Khoraminia[2], J. Silva-Rodríguez[3], U. Kiraz[4,5], G. J. L. H. van Leenders[6], T. Eftestøl[1], V. Naranjo[7], E. A. M. Janssen[4,5], T. C. M. Zuiverloon[2], K. Engan[1]

[1] Dept. of Electrical Engineering and Computer Science, University of Stavanger, 4021 Stavanger, Norway

[2] Dept. of Urology, Erasmus MC Cancer Institute, University Medical Center, 3015 GD Rotterdam, The Netherlands

[3] ÉTS Montréal, Montréal, Québec 1011, Canada

[4] Dept. of Pathology, Stavanger University Hospital, 4011 Stavanger, Norway

[5] Dept. of Chemistry, Bioscience and Environmental Engineering, University of Stavanger, 4021 Stavanger, Norway

[6] Dept. of Pathology and Clinical Bioinformatics, Erasmus MC Cancer Institute, University Medical Center, 3015 GD Rotterdam, The Netherlands

[7] Instituto de Investigación e Innovación en Bioingeniería, I3B, Universitat Politècnica de València, Valencia 46022, Spain

Under review. Preprint available on arXiv:

https://arxiv.org/abs/2405.15264

Abstract:
We present a pioneering investigation into the application of deep learning techniques to analyze histopathological images for addressing the substantial challenge of automated prognostic prediction. Prognostic prediction poses a unique challenge as the ground truth labels are inherently weak, and the model must anticipate future events that are not directly observable in the image. To address this challenge, we propose a novel three-part framework comprising of a convolutional network based tissue segmentation algorithm for region of interest delineation, a contrastive learning module for feature extraction, and a nested multiple instance learning classification module. Our study explores the significance of various regions of interest within the histopathological slides and exploits diverse learning scenarios. The pipeline is initially validated on artificially generated data and a simpler diagnostic task. Transitioning to prognostic prediction, tasks become more challenging. Employing bladder cancer as use case, our best models yield an AUC of 0.721 and 0.678 for recurrence and treatment outcome prediction respectively.

14.1 Introduction

The introduction of digital pathology, characterized by the digitization of tissue sections into whole slide images (WSI) through microscopy scanners, has opened up numerous possibilities. A prevailing trend in computational pathology (CPATH) research involves utilizing image processing and machine learning to develop tools that assist pathologists in visualization, region of interest (ROI) extraction, and diagnostic tasks [123].

Prognostic prediction is generally acknowledged as more challenging than diagnostic prediction as it involves forecasting future events and outcomes. Prognosis is arduous due to histological diversity, observer variability, and tumor heterogeneity [177]. Additionally, diverse treatment responses and recurrence patterns must be considered. Urinary bladder prognostic prediction exemplifies this complexity as a result of diverse treatment responses, recurrence patterns, and the absence of distinct markers and varied clinical factors [10].

Plenty of research in CPATH is dedicated to diagnostic prediction, often relying on manually selected ROIs during both model learning and inference stages. For prognostics, weakly labeled data is commonly employed, with patient-based labels defining treatment outcomes, recurrence, or disease progression. Consequently, there is no guarantee that a particular region contains the necessary information, or that the essential information is genuinely present in the WSI.

In this study, we present an automated deep learning pipeline for prognostic predictions from WSI, trained and tested on weakly labeled data. A nested multiple instance learning method with attention mechanisms (NMIA), recently proposed by our research group, is used for the first time in an end-to-end problem [250]. The pipeline utilizes extensive unannotated regions to enhance feature representations and explores the impact of selectively choosing informative areas within the slides. This work represents a pioneering investigation into the application of prognostic predictions in urinary bladder cancer. The main contributions of this paper are summarized as:

(i) Introduction of an automated deep learning pipeline for prognostic predictions from WSI, leveraging weakly labeled data. The pipeline incorporates NMIA, a novel approach applied for the first time in an end-to-end problem.

(ii) Utilization of extensive unannotated regions to enrich feature representations and investigation into the impact of selectively choosing informative areas within the slides guided by domain knowledge, marking a pioneering exploration into the application of prognostic predictions in urinary bladder cancer.

14.2 Background & Related Work

14.2.1 Urinary Bladder Cancer

Non-muscle invasive bladder cancer (NMIBC) conform approximately 75% of bladder cancer diagnoses [109]. The 2022 version of the European Association of Urology (EAU) guidelines on NMIBC suggests that patients are stratified into risk groups based on the hazard to progress to a muscle-invasive disease [10]. The hazard score depends on significant clinical and pathological factors, which are themselves time-consuming to determine and frequently result in variations among uropathologists. Currently, intravesical Bacillus Calmette–Guérin (BCG) is the gold standard adjuvant therapy for high-risk non-muscle invasive bladder cancer (HR-NMIBC). Unfortunately, BCG treatment causes many severe adverse reactions, and up to 50% of the patients will develop BCG resistance, thus resulting in recurrence or progression [11]. Identifying these patients at the first stage would significantly reduce the recurrence rate and treatment cost, hence contributing to more adequate patient-based treatment strategies. Nonetheless, bladder WSI are sizable and often contain disorganized, fragmented tissue sections, with a significant number of artifacts and other non-diagnostically-relevant tissue [109]. The unique challenges of urinary bladder cancer WSI have served as inspiration for the three-step pipeline presented in this study.

14.2.2 Data Modalities

Computer-aided diagnosis (CAD) systems that utilize machine learning techniques for medical imaging analysis of WSI have shown effective ways to reduce subjectivity and speed up the diagnostic process [123]. WSI are pre-stored at various magnification levels, allowing pathologists to quickly adjust the zoom level, analogous to physical microscopes. Lower magnification is typically used to view tissue-level morphology, while higher magnification is useful for examining cell-level features. In CPATH, imaging

techniques that rely on convolutional neural networks (CNNs) are considered the best option for extracting features from histological images [118]. However, prognostic applications have primarily relied on clinicopathological information more than on image data [180]. Both clinicopathological information and image features are derived from the visual characteristics of tumor tissue. Image features are extracted directly from the image, while clinicopathological information depends on external observations. Nonetheless, recent deep learning prognostic applications employ image features, while providing reasonably accurate prognostic predictions. Although there are methods based on clinicopathological or image data alone, hybrid solutions have also been proposed with state-of-the-art performance [37, 181]. Another promising type of data is that of genomics, specifically next-generation sequencing, which has altered the understanding and assessment of cancer[260]. However, next-generation sequencing is an expensive technology, still in its early implementation phase, making it an inaccessible solution for many pathology laboratories at present. In contrast to this, hematoxylin and eosin-stained (H&E) WSI provide an affordable solution and practical choice for routine diagnostic and research purposes. Therefore, we explore H&E WSI both with and without clinical information in this study.

14.2.3 Non-Supervised Learning Methods

One challenge in CPATH is the lack of WSI with detailed or region of interest (ROI) annotations, which is often expensive and time-consuming to acquire. To address this, weakly supervised methods have been proposed for training deep learning models on WSI [123, 245, 261–263]. Weakly supervised methods emerge as an advantageous approach for prognostic applications as they accommodate the uncertainty and heterogeneity of medical data. Among the diverse weakly supervised methods, attention-based multiple instance learning (AbMIL) is a popular approach [84, 199, 264, 265]. The method uses an attention mechanism to selectively focus on regions of interest within an image, allowing the model to learn from weakly labeled data. One diagnostic application using weakly supervised deep learning on WSI is that of predicting the pathological grade of the patient [31, 33, 253]. Such approach was able to achieve performance on par with supervised methods, while reducing the amount of annotated data required. Cancer survival prediction using WSI was performed with AbMIL, as demonstrated in [266–270]. It is highlighted that the attention-based approach improved

the performance of the model. However, WSI present scattered tissue, with countless instances. Hence, a straightforward MIL approach may not be suitable, as WSI could be densely populated with noisy instances. This led us to propose NMIA, which restricts cross-contamination among regions within the images [250].

The relationship between image features and patient outcomes can be complex and difficult to discern. To address this lack of correlation, self-supervised methods can overcome the disparity. In recent years, contrastive learning has emerged as a promising technique for learning feature representation from large unlabeled datasets. Contrastive learning is a type of learning with the aim of training a feature extractor, using a contrastive loss function [271]. This learning approach has the capacity to utilize extensive unannotated regions for enhancing feature representation. Typically, a contrastive module serves as a preliminary step before classification, facilitating the extraction of feature representations [272, 273]. Moreover, the acquisition of transformation-independent features has proven effective in mitigating stain variation [274, 275]. It has also been employed for maximizing feature similarity between areas from the same WSI [75, 276]. Contrastive learning in WSI prognostics is mostly unexplored [277]. We adopt SimCLR and variations for exploiting underlying prognostic patterns in WSI [271].

14.3 Dataset

In this study, we have gathered NMIBC WSI from two distinct cohorts. The total number of patients per application is displayed in Table 14.1.

Set	S_{EMC}		S_{SUH}	
	BCG-R	BCG-NR	Rec	NoRec
Train	272 (72)	81 (42)	113 (0)	107 (0)
Validation	25 (24)	25 (22)	18 (0)	12 (0)
Test	25 (25)	25 (25)	27 (0)	23 (0)

Table 14.1: Description of the patient sets, S_{EMC} and S_{SUH}, and how many patients in each group divided in train/val/test splits. The number between parenthesis indicates the number of patients with annotated WSI.

We have available HR-NMIBC WSI from a multi-centre cohort provided by Erasmus Medical Center (EMC), Rotterdam, The Netherlands. We

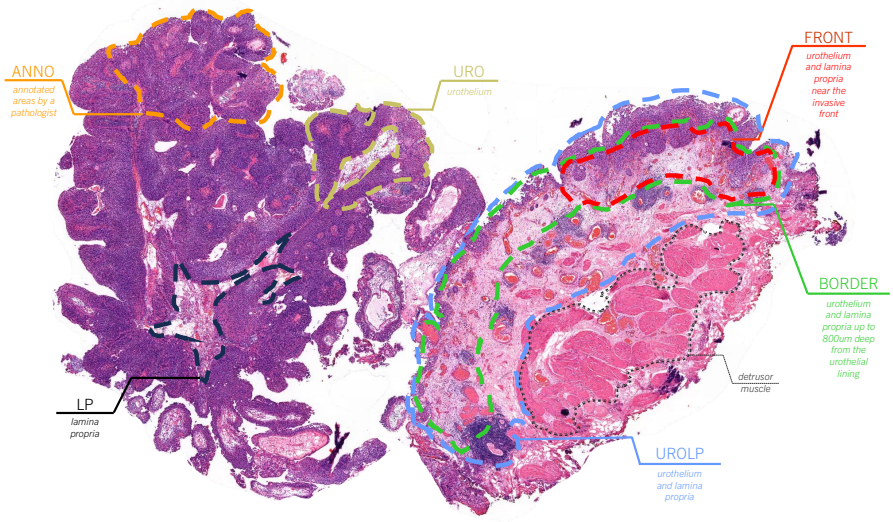


Figure 14.1: Overview of ROI generation. The process of extracting ROIs from raw WSI involved either a tissue segmentation algorithm and/or pathologist's annotations. The annotations highlighted areas that were deemed prognostically significant for predicting outcome, while the algorithm provided masks that highlighted different tissue types. Subsets of tissue were extracted using urothelium and lamina propria. For magnification levels, the study explored two mono-scale approaches using 10x and 20x, as well as a multi-scale method using three magnifications (2.5x, 10x, 40x).

denote this dataset S_{EMC}, and use it for BCG response prediction. Let BCG-R and BCG-NR denote BCG responder and non-responder tumor, correspondingly. BCG-NR corresponds to BCG failure according to EAU guidelines, excluding BCG intolerance. A total of 453 patients and 503 WSI formed the dataset. Since the treatment outcome is related to the patient as a whole and not to a specific region of the tumor, patches from various WSI that belong to the same patient were merged as a single entity. Not all slides contained annotations, and among those that were annotated, none of the WSI were fully annotated due to time and database storage constraints. Annotations contain tissue types, artifacts, grading and staging. Moreover, a detailed report of clinicopathological information per patient was disclosed with information about BCG treatment guidelines, pathological diagnoses, and patient demographics. We utilized clinical variables of gender, age, smoking status, grade, stage, concomitant carcinoma in situ, size and focality of the tumor.

We also included a total of 300 WSI corresponding to 300 different

NMIBC patients from Stavanger University Hospital (SUH), Stavanger, Norway. We denote this dataset S_{SUH}, and use it for recurrence prediction. Let NoRec and Rec denote no recurrence and recurrence, respectively. Rec was defined as recurrent tumors in the bladder only, iwth a median follow-up of 82 months. No WSI were annotated with ROI, but weak labels regarding recurrence outcome were available.

With regards to ROI definition, two main strategies were employed: using annotated areas or areas defined from an automatic tissue segmentation algorithm. Fig. 14.1 highlights the various ROIs explored in this study. The automatic definition of ROIs is later described in 14.4.1. Concerning annotations, an engineer with specific training in bladder pathology (F. K.) partially annotated 217 of the total 503 EMC patient slides from S_{EMC}, under the supervision of an experienced uropathologist (G. vL.). This subset of data is referred as D_{ANNO}. The annotation process consisted of general tissue type annotations, and in some cases, sub classes indicating grading, presence of tumor infiltrating lymphocytes, flat lesions, and invasive areas. First, a coarse annotation of the WSI was done, identifying and labeling the main regions of interest within the images. Then, a quality control process was implemented to ensure the annotation's quality and consistency by getting external revision from expert uropathologists.

14.4 Methods

In the upcoming subsections, we introduce a three-step fully-automated pipeline for WSI prognosis that combines region of interest (ROI) extraction, contrastive learning for feature representation and multiple instance learning (MIL) for predicting the concluding prognostic outcome, as depicted in Fig 14.2. This approach enables us to optimize the model by leveraging the benefits of these techniques. It ensures the inclusion of important instances for predicting clinical outcome from WSI visual cues, while maintaining computational feasibility. Ultimately, the proposed steps for prognostic predictions in histopathological imaging are the following:

A Define and extract ROIs using a tissue segmentation algorithm for tile extraction strategies.

B Train a feature extractor G_θ to generate an intermediate dataset $\mathcal{H}$ using contrastive learning.

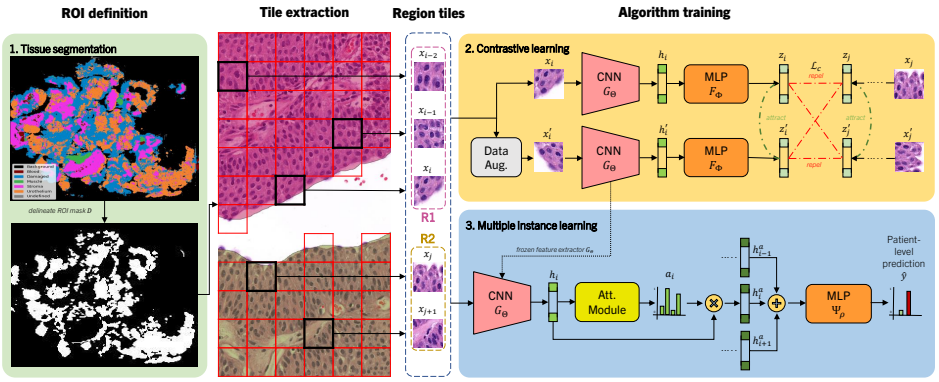


Figure 14.2: Deep learning pipeline for prognostic outcome prediction. 1) A tissue segmentation is employed for delineating a ROI of choice D. Then,tiles are extracted from WSI regions for training an algorithm. 2) Contrastive learning is employed to learn representations of the tiles. 3) The representations are then used to train an AbMIL model that predicts the prognostic outcome. This approach compresses an end-to-end pipeline where the raw input image data is broken down and processed for predicting clinical outcome.

C Use image feature embeddings $\mathcal{H}$ for prognostic classification using MIL.

14.4.1 Automatic Region of Interest Segmentation

The current study will explore various ROI configurations as the localization of the tissue of interest is unknown. While annotations can be expensive and inflexible, automatic tissue segmentation algorithms provide the flexibility to redefine ROIs based on different clinical considerations. A tissue segmentation algorithm was proposed in [26] for S_{SUH}, while an active learning-based approach for tissue segmentation was developed in [254] for S_{EMC}. Both models share the same architecture, utilizing a tri-scale CNN backbone that leverages different magnifications for each input CNN. The tissue segmentation algorithms work at patch level, and classify all patches x in the WSI as $y \in \mathcal{Y} = \{$urothelium, lamina propria, muscle, blood, damage, background$\}$. The dataset resulting from extracting patches with label y is denoted D_y. For example, urothelium tissue is the most prominent source of information in urothelial bladder carcinoma, and the dataset of patches extracted from these regions are denoted D_{URO}. In conjunction with urothelium, lamina propria may serve an important role for influencing

the growth of the tumor [182]. The dataset of patches extracted from lamina propria is denoted D_{LP}, and the union urothelium and lamina propria $D_{\text{UROLP}} = D_{\text{URO}} \cup D_{\text{LP}}$. However, these D_y are solely defined based on tissue types, and to meticulously analyze and concentrate on the pertinent area of interest for comprehending the disease's status, it is imperative to define tailored ROIs. Consequently, exploiting domain knowledge through segmentation maps is a pivotal aspect of this work. Notably, not all lamina propria might be interacting with the tumor. Therefore, we defined a depth of $800\mu m$ based on medical knowledge for defining possible tumor and immune response interactions based on medical knowledge. The union of the boundary between urothelium and lamina propria defines the dataset $D_{\text{BORDER}} \subset D_{\text{UROLP}}$. This is accomplished by applying a disk dilation operation on the urothelium and lamina propria masks, where the disk radius is determined by pixel size, and subsequently segment the overlapping area to extract the bordering region. Fig. 14.3 displays an schematic representation of the self-defined ROI D_{BORDER}. Furthermore, the invasive front of the tumor, which represents the most aggressive part of the tumor, could potentially provide the most significant features for comprehending the current state of the tumor in relation to the patient's immune system [183]. Therefore, we refine D_{BORDER} using a region-growing algorithm along these borders to exclude areas lacking muscle tissue within a tissue section, thus defining $D_{\text{FRONT}} \subset D_{\text{BORDER}}$. To exclude distant muscle areas from consideration, we apply the same distance threshold of $800\mu m$, thus ensuring the focus remains on the invasive front regions.

Regarding tile extraction strategies, various magnification levels and tile sizes are used. In the case of mono-scale models, we used a tile size of 256×256 for 10x magnification and 512×512 for 20x magnification. This decision was made to ensure that the patches covered the same physical area, i.e. field of view. In the case of multi-scale models TRI, we used a tile size of 128×128 for all three magnification levels (40x, 10x, and 2.5x), extracted following the approach described in [120].

14.4.2 Feature Extraction via Contrastive Learning

Contrastive learning is a specific approach within metric learning that focuses on comparing similar and dissimilar pairs of examples, based on a siamese structure, by attracting and repelling representations accordingly [271]. A contrastive model is trained as follows: given a batch of N random samples from a training set of X images, a set of two image

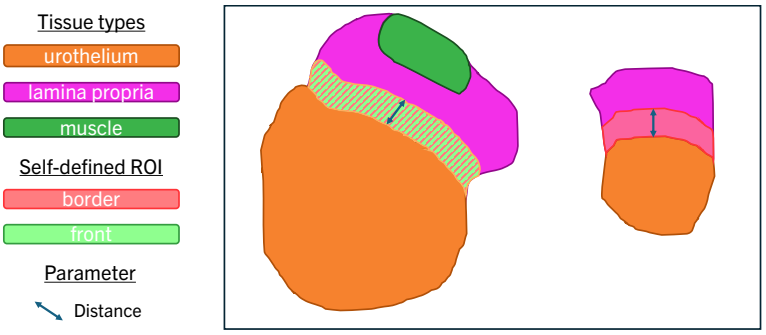


Figure 14.3: Schematic representation of self-defined ROI generation. Guided by the segmentation mask of various tissue types, we apply diverse image processing morphological operations to define ROIs D_y based on domain knowledge from expert pathologists. In the example displayed, we enlarge the urothelium and lamina propria masks applying dilation, limited by a distance parameter determined on clinical expertise. The resulting overlapping areas represent D_{BORDER}. We further delineate a subset of D_{BORDER} by extracting only those areas where muscle is present within the same tissue section, aiming to represent the potential invasive front in D_{FRONT}.

transformations are applied to all images in the batch in order to obtain $2N$ augmentations. These transformed images are forwarded though a feature extractor $G_\theta : \mathcal{X} \to \mathcal{H}$ and projected into a low-dimensional feature space through a multi-layer perceptron (MLP) $F_\phi : \mathcal{H} \to \mathcal{Z}$, to be later l_2 normalized. Let $\text{sim}(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}_i^\top \mathbf{z}_j / \|\mathbf{z}_i\| \|\mathbf{z}_j\|$ be the cosine similarity between l_2 normalized vectors $\mathbf{z}_i$ and $\mathbf{z}_j$, then the loss function for a positive pair is defined as:

$$\mathcal{L}_c = -\frac{1}{2N} \sum_{i \in I} \log \frac{\exp(\text{sim}(\mathbf{z}_\text{i}, \mathbf{z}_\text{i}')/\tau)}{\sum_{j=1}^{2N} 1_{[j \neq i]} \exp(\text{sim}(\mathbf{z}_\text{i}, \mathbf{z}_\text{j})/\tau)} \tag{14.1}$$

where $\mathbf{z}_i'$ is the augmented representation of $\mathbf{z}_i$, $1_{[j \neq i]}$ is a binary indicator to indicate all other instances than i, and τ is a temperature coefficient to control the strenght of penalties on hard negative samples. The contrastive loss has the objective of ensuring the model learns strong feature representations regardless of the augmentation applied, thus increasing the robustness to variability in the input image. The contrastive loss rewards the model for creating similar features for both augmentations from the same image, while increasing feature dissimilarity between other augmentations from images in the batch. $\mathcal{L}_c$ corresponds to the unsupervised version, although the supervised contrastive learning loss $\mathcal{L}_{sc}$ does exist [278]. The supervised loss $\mathcal{L}_{sc}$ considers images from the same class in the batch, using the

corresponding image label y, and does not punish the model for generating similar representations among images from the same class:

$$\mathcal{L}_{sc} = \sum_{i \in I} \frac{-1}{|P_{(i)}|} \sum_{p \in P_{(i)}} \log \frac{\exp(\text{sim}(\mathbf{z}_\text{i}, \mathbf{z}_\text{i}')/\tau)}{\sum_{j=1}^{2N} 1_{[j \neq i]} \exp(\text{sim}(\mathbf{z}_\text{i}, \mathbf{z}_\text{j})/\tau)} \tag{14.2}$$

where $P_{(i)} \equiv p : y_p = y_i$, while $|P_{(i)}|$ represents its cardinality. $\mathcal{L}_{sc}$ is preferred when labeled data is available, allowing for explicit learning of relevant representations and improved model performance. However, when labeled data is scarce or difficult to obtain, $\mathcal{L}_c$ can be a practical option.

We further explore the implementation of contrastive learning defining a multi-task learning loss. For multi-task learning, two separate projection heads for both $\mathcal{L}_c$ and cross entropy loss $\mathcal{L}_{ce}$ are simultaneously trained. $\mathcal{L}_{ce}$ is calculated using the output predictions of a classifier module $C_\eta : \mathcal{H} \rightarrow \hat{\mathcal{Y}}$. As labels are a requirement for calculating $\mathcal{L}_{ce}$, D_{ANNO} is employed. For computing the loss for multi-task contrastive learning, we combine the loss from two separate projections:

$$\mathcal{L}_{multi} = \alpha_c \mathcal{L}_c + \alpha_{ce} \mathcal{L}_{ce} \tag{14.3}$$

where α_c and α_{ce} are the scaling factors for the unsupervised contrastive and supervised cross entropy losses, respectively.

14.4.3 Prognostic Outcome Classification via Multiple Instance Learning

Multiple instance learning (MIL) is a suitable approach for prognostic classification due to its ability to handle inherent uncertainties, diversity and intricacies present in medical data [248]. A dataset $\mathcal{H}, \mathcal{Y} = \{(\mathbf{H}^i, y^i), \forall i = 1, ..., N\}$ is formed of pairs of bag instances $\mathbf{H}$ and their corresponding labels y, where i denotes the current sample for a total of N samples. A bag $\mathbf{H}$ consists of instances $\mathbf{h}_l$:

$$\mathbf{H} = \{\mathbf{h}_l, \forall l = 1, ..., L\} \tag{14.4}$$

where L is the number of instances in the bag. Among MIL model architecture variants, we adopted attention-based multiple instance learning (AbMIL). Given a label for a patient, a model should infer which ROIs

visual features lead to predicting the patient's prognostic outcome. An attention score a_i for a feature embedding $\mathbf{h}_i$ can be calculated as:

$$a_i = \frac{\exp\{\mathbf{w}^\top(\tanh(\mathbf{V}\mathbf{h}_\mathbf{i}^\top) \odot \text{sigm}(\mathbf{U}\mathbf{h}_\mathbf{i}^\top))\}}{\sum_{l=1}^{L}\exp\{\mathbf{w}^\top(\tanh(\mathbf{V}\mathbf{h}_\mathbf{l}^\top) \odot \text{sigm}(\mathbf{U}\mathbf{h}_\mathbf{l}^\top))\}} \tag{14.5}$$

where $\mathbf{w} \in \mathbb{R}^{L\times 1}$, $\mathbf{V} \in \mathbb{R}^{L\times M}$ and $\mathbf{U} \in \mathbb{R}^{L\times M}$ are trainable parameters and $\odot$ is an element-wise multiplication. Furthermore, the hyperbolic tangent $\tanh(\cdot)$ and sigmoid $\text{sigm}(\cdot)$ are included to introduce non-linearity for learning complex applications. The strength of the attention modules is not only in terms of interpretability, but also in predictive power, as attention scores are directly influencing the forward propagation of the model. Once the attention scores are obtained, we obtain the patient prediction $\hat{y}$ as:

$$\hat{y} = \Psi_\rho(\mathbf{A} \cdot \mathbf{H}) \tag{14.6}$$

where Ψ_ρ is a MLP acting as a patient classifier. Additionally, we propose using NMIA [250]. NMIA defines a bag can consisting of multiple sub-bags, which contain the instances themselves. This serves to further stratify into clusters or regions, and accurately represent the arrangement of the scattered data, where tiles belong to particular tissue areas and these themselves to the WSI. A bag-of-bags for a WSI $\mathbf{H}_{\text{WSI}}$ contains a set of inner-bags, or regions, $\mathbf{H}_{\text{REG},k}$:

$$\mathbf{H}_{\text{WSI}} = \{\mathbf{H}_{\text{REG},k}, \forall k = 1, ..., K\} \tag{14.7}$$

where the number of inner-bags K varies between different WSI. Ultimately, $\mathbf{H}_{\text{REG},k}$ contain instance-level representations $\mathbf{h}_{\text{TILE},l}$ of tiles located within the physical region.

For this project, we explored three configurations: using image data, using clinicopathological data and fusing both. For image data, we applied a weakly supervised architecture. For clinical data, we sorted it into a 1-dimensional vector $\mathbf{h}_{var}$ and fed it directly to the patient classifier Ψ_ρ. As for the combination of both, we used a weakly supervised architecture for generating the patient embedding representation and concatenated the clinical features to said embedding as $\mathbf{h}_{cli} = [\mathbf{A} \cdot \mathbf{H}, \mathbf{h}_{var}]$.

14.5 Experimental Setup

In this section, we present the carefully designed experiments. First, to evaluate the proposed pipeline structure, we present experiments on artificial data as well as an application with region based labels. From there on, we consider two prognostic applications: BCG response prediction and recurrence prediction. Hyperparameters, like the choice of optimizer, loss, and others are identical in all experiments. The code is available at https://github.com/Biomedical-Data-Analysis-Laboratory/HistoPrognostics.

ROI Extraction

The fully automatic tissue segmentation was performed according to [26] for recurrence S_{SUH} and [254] for BCG treatment S_{EMC}. Different regions were extracted as explained in Section 14.4.1.

Contrastive Feature Representations

We define a temperature parameter to 0.07 for the calculation of the loss for contrastive learning, as explained in Section 14.4.2. We experiment with different CNN backbones; VGG16, DenseNet121 and ResNet18, with initial weights θ_I, from pretraining on ImageNet [256]. For the supervised approaches, labels of grading and presence of TILs were used as diagnostic factors. In order to weigh the impact of the training size, we also run experiments of the unsupervised variant with larger, but limited, training samples. With respect to multi-task learning of unsupervised contrastive and cross entropy classification, we set the parameters α_c and α_{ce} to 1.0 and 0.5, respectively. Adam was set as optimizer with a learning rate of 1e-4, a batch size of 128, and a total of fixed 10 epochs. The augmentations applied consisted of flip, flop, rotation, affine transformations, and color jittering.

Prognosis Classification

Bladder cancer recurrence is indeed regarded as a manifestation of treatment failure, as it indicates that the initial treatment did not effectively eradicate all cancer cells. Building upon this rationale, we will employ the data set S_{EMC}, see Table 14.1, in our decision-making process regarding the selection of feature extractors, contrastive loss functions, ROI selection,

and magnification levels for both prognostic applications. This choice derives from the larger number of patients in the dataset, being a more representative sample of the population under study. Focal Tversky loss (FTL) is employed [257]. FTL employs two parameters, denoted as α_l and γ_l, which allow for adjusting the focus on different classes and handling the difficulty of training examples, respectfully. We also set an early stopping criteria of 30 epochs based on the AUC score on the validation set. We also implement a 5-runs Montecarlo with a 5% dropout for sampling purposes. A grid hyperparameter search is done to find the optimal values for the given task. The search includes bag sampling n_b, learning rate lr, optimizer opt, dropout rate d_r, number of neurons in the classifier n_{Θ_ρ} and attention mechanism n_{att}, loss functions parameters α_l and γ_l. The list of hyperparameters with their corresponding possible values and the resulting choice can be seen in Table 14.2:

Hyperparameter	**List of values**
lr	$10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}$
opt	SGD, Adam
n_b	4, 16, 64, 256, L
d_r	0.1, 0.2, 0.3, 0.4, 0.5
α_l	0.0, 0.3, 0.6, 0.9
γ_l	0.5, 1, 2
n_{Θ_ρ}	128, 512, 1024, 4096
n_{att}	128, 512, 1024, 4096
Hyperparameter choice	$lr = 10^{-2}, opt$ =SGD, $n_b = 64$ $d_r = 0.2, \alpha_l = 0.9, \gamma_l = 2.0,$ $n_{\Theta_\rho} = \{4096, 2048\}, n_{att} = 4096$

Table 14.2: Results of hyperparameter search for optimizing predictive performance.

14.6 Preliminary Experimentation

The three-step pipeline is convoluted, and prognostic labels are weak, typically involving one label per patient. To evaluate the pipeline, we conduct experiments in more controlled settings. Firstly, we aim to assess the pipeline using synthetic data generated from distributions with varying

degrees of overlap. Thereafter, we evaluate the pipeline on actual WSI data, focusing on a diagnostic task with strong region-based labels, specifically the detection of regions containing lymphocytes.

14.6.1 Effects of Different Data Distributions

We define three different matching distributions pairs, which can be described as distributions with minor, partial overlap, or significant overlap. The degree of overlap is tuned using the mean μ and standard deviation σ parameters. For all experiments, we define $\mathcal{P}_{3,2.5}$ as our class 0. For the matching distribution of class 1, we use $\mathcal{P}_{-3,1}$, $\mathcal{P}_{0,2}$ and $\mathcal{P}_{2,1.5}$ for minor, partial and significant overlap, respectively. A total of 150 bags are created, balanced among two classes, with a 90, 30, 30 split for train, validation and test, respectively. The number of instances per bag is variable $N \in [3000, 7000]$, without replacement. The number of positive instances is limited to be less than half.

Results indicate that it is uncomplicated to discern between $\mathcal{P}_{-3,1}$ and $\mathcal{P}_{0,2}$ distributions, as shown in Table 14.3. However, we observe that for an overlapping distribution $\mathcal{P}_{2,1.5}$ the learning breaks down. We wanted to go further into finding the breaking point for the partial overlap distribution $\mathcal{P}_{0,2}$, and as such, we tried different parameters regarding the bag label balance in the train set and the representation of positive samples in bags. We look into a percentage of positive bags of 25 and 40%, as well as the number of positive samples in the range of 0 to 50%, and 0 to 75%. Results reveal the significance of class imbalance in bag classification, and the representation of positive and negative samples within the bags. The

Class 1	**AUC**
$\mathcal{P}_{-3,1}$	**1.000**
$\mathcal{P}_{0,2}$	**1.000**
$\mathcal{P}_{2,1.5}$	0.500

Table 14.3: Prediction probabilities for discerning $\mathcal{P}_{3,2.5}$.

MIL	**Weights θ**	**AUC**
MIL_{1+}	θ_I	0.833
	θ_{CE}	0.938
	θ_{SC}	0.909
	θ_{C}	0.940
	θ_{MULTI}	0.938
MIL_{t+}	θ_{C}	**1.000**

Table 14.4: TIL detection comparison for different feature extractors.

classification of highly imbalanced bag datasets can be challenging. While augmenting the number of positive instances can enhance the instance classification performance, it is insufficient to achieve satisfactory bag classification results. The performance naturally improves as the bag distribution becomes more balanced, but it is not enough unless the positive class is over-represented.

14.6.2 Detection of Lymphocytes

We evaluated the performance of our proposed solution in a diagnostic task using the same WSI as those used in S_{EMC}. Specifically, we defined a problem of detecting the presence of tumor-infiltrating lymphocytes (TILs), immune cells which have been associated with improved patient outcomes [11]. Tiles of size 256 $\times$ 256 were extracted from annotated lamina propria areas, with and without TILs, at 40x magnification. As a pre-processing step, tiles were normalized and resized to 224 $\times$ 224. Augmentations performed consisted of rotation and jittering operations. We define two experiments for different percentages of tiles with stromal TILs. The first one, MIL_{1+}, follows the standard MIL convention of detecting at least one tile with TILs. As for the second one, MIL_{t+}, we define a threshold t for which more than 50% of the tiles must contain TILs.

We employed MIL_{1+} to determine the optimal configuration for feature extraction. The results are presented in Table 14.4. Pre-trained features from natural images indeed capture relevant features for histopathological classification. However, the unsupervised approach θ_{C} exhibits superior performance in bag-level classification. MIL_{t+} is therefore only tested with G_{θ_C}, and this provides the best result with an AUC of 1.0. This might be due to the fact that labels are noisy, as many of the tiles annotated TIL-free actually contain some, even if the count is low. Using MIL_{t+} makes it easier for the model to find a threshold t than an absolute presence or absence given the noisy weak labels. Consistent with the observations on synthetic data, bags containing a substantial number of positive instances exhibit superior performance. This implies that accurate predictions rely on the existence of several positive instances within a bag, with less emphasis on the presence of positive instance outliers.

14.7 Prognostic Experiments

In this section, we utilize the proposed three-step pipeline for two prognostic applications: BCG response prediction on the S_{EMC} dataset, and recurrence prediction on the S_{SUH} dataset. Given the numerous choices for contrastive learning loss, backbone, ROI, and magnification levels, coupled with the computational intensity of learning and inference on gigapixel images, we adopt a systematic experiment approach. Therefore, we focus on testing one factor at a time and restrict further testing to the most promising results. In order to identify the optimal choices, we couple an AbMIL classification module to assess the resulting classification performance, using the validation subset. In subsections 14.7.1-14.7.6, we use BCG response prediction with S_{EMC} as the reference task, while recurrence prediction with S_{SUH} is discussed in subsection 14.7.6. Finally, subsection 14.7.8 discusses the interpretability of trained models.

14.7.1 Feature Extraction and Contrastive Learning

We aim to select a CNN backbone of preference for the feature extractor G_θ, referring to the second step in the pipeline. The most commonly used backbones in CPATH literature are DenseNet, ResNet and VGG [118]. To ensure a fair performance comparison in a contrastive learning approach, we will incorporate the use of different labels, which will either be unsupervised, supervised contrastive or multi-task learning. In this experiment, we constrict the ROI to be D^{20x}_{URO}, with patches extracted at 20x. The urothelium tissue type is widely recognized as highly informative, and the choice of 20x magnification strikes a balance between capturing morphological structure context and preserving cellular details.

Weights θ	**DenseNet121**	**ResNet18**	**VGG16**
θ_I	0.576(0.029)	0.474(0.009)	0.549(0.005)
θ_{C}	**0.672(0.032)**	0.628(0.017)	0.506(0.022)
θ_{SC}	0.521(0.054)	0.568(0.009)	0.480(0.036)
θ_{MULTI}	0.515(0.031)	0.506(0.042)	0.434(0.048)

Table 14.5: Validation AUC scores for D^{20x}_{URO} BCG. We compare various CNN architectures and feature extraction strategies. The results show the mean and standard deviation over 5 runs.

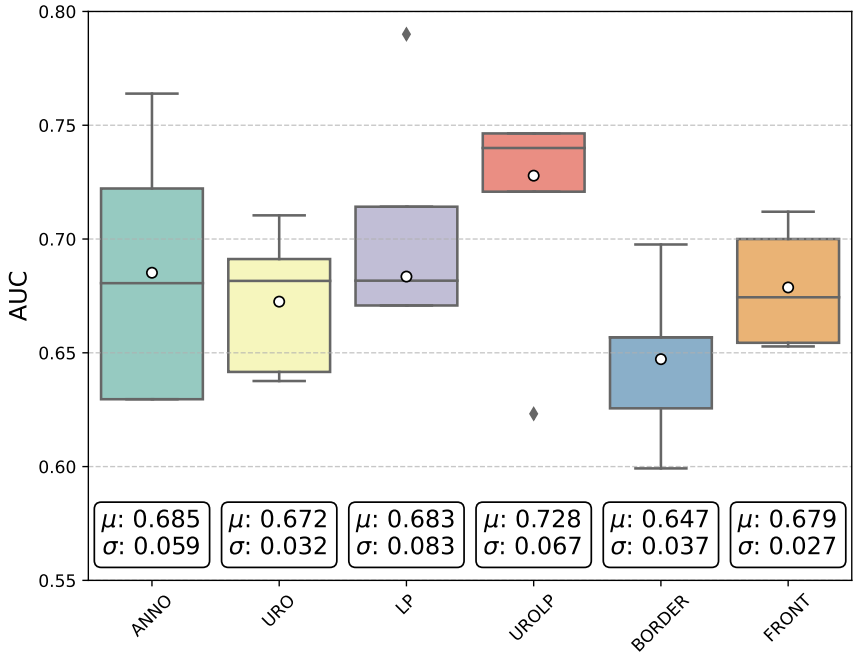


Figure 14.4: Box plot illustrating AUC performance variation across validation sets with different ROIs for 20x magnification. Notably, D^{20x}_{UROLP} emerges as the top performer amongst the ROIs. White dots represent the average value μ, and black diamonds represent outliers. The results show the mean μ and standard deviation σ over 5 runs.

Results for classification can be found in Table 14.5. DenseNet121 offers promising results in terms of AUC performance for classification, without compromising computational efficiency. Therefore, we will use DenseNet121 as the backbone for the remaining experiments. The best performing contrastive learning strategy is unsupervised $\mathcal{L}_c$. Hence, we will proceed to

Magnification	UROLP	FRONT	ANNO*
10x	0.621(0.021)	0.647(0.033)	**0.790(0.113)**
20x	**0.728(0.067)**	0.679(0.027)	0.685(0.059)
TRI	0.649(0.014)	0.621(0.038)	0.615(0.029)

Table 14.6: Validation AUC scores for BCG at different ROIs and magnification levels. The results show the mean and standard deviation over 5 runs.
* Not all train and validation WSI were annotated.

use the frozen weights from the unsupervised method θ_{C}.

14.7.2 Region of Interest Selection

After determining the CNN backbone and contrastive learning strategy, the next step involves identifying the ROI with the best discriminative capability for the prognostic task. As highlighted in Fig. 14.1, the ROIs are found from either manual annotations or from the output of the automated tissue segmentation model. From the automatically segmented regions, we obtain D^{20x}_{URO}, D^{20x}_{LP}, D^{20x}_{UROLP}, $D^{20x}_{\mathrm{BORDER}}$, D^{20x}_{FRONT}, while D^{20x}_{ANNO} is generated from the annotated set.

A comparison is shown in Fig 14.4. The comparison underscores D^{20x}_{UROLP}'s advantage, showcasing its consistent and balanced classification performance with an average AUC score of 0.728. That said, D^{20x}_{FRONT} has the lowest variance across runs, reaching the most consistent results. D^{20x}_{ANNO} demonstrates the second-highest average AUC score, but the highest variance. This implies a promising characteristic for future development of diagnostic models. Emulating a pathologist sense of expertise, a model could automatically identify diagnostically relevant areas and generate annotations. Then, these generated annotations could be used for further enhancing feature discrimination for subsequent prognostic models. That said, in our investigation, only a limited set of tiles is annotated. This results in a mismatch in training size between datasets automatically generated and annotations. Moreover, an independent system cannot rely on expert input during the inference stage. Moving forward, we take all three aforementioned ROIs for determining the optimal magnification input choice.

14.7.3 Magnification Level Choice

We aim to assess the influence of varying magnification levels on the performance, testing mono-scale magnification levels of 10x, 20x and multi-scale TRI (2.5x, 10x, 40x). The results are shown in Table 14.6, and allows us to assess the model's accuracy in capturing both broader tissue context and finer cellular details, as well as contextual neighbouring regions as per the multi-scale input.

Utilizing a fixed magnification level throughout the image consistently yields AUC levels that hover around 0.7 at 20x magnification, the highest for the automatic ROI D^{20x}_{UROLP}. However, it's worth noting that the predictive performance for 10x is worse. For comparison, D^{10x}_{ANNO} obtained on average

higher AUC values, yet, constrained to expert input and lack of a complete set of annotations. Despite of its capacity to capture structural intricacies, the multi-scale model TRI stands with lower performance. TRI performs worse possibly because the larger input requires more training data for generalization.

14.7.4 Weakly Supervised Aggregation for Treatment Outcome Prediction

We conduct a comparison between five aggregation techniques in weakly supervised learning. These include majority voting, max, mean, AbMIL and NMIA. These techniques are evaluated to determine their effectiveness in combining and summarizing information from multiple instances.

In Table 14.7, we present the obtained results. Among these techniques, those involving majority voting, mean and max aggregation, which do not include in-built attention mechanisms, demonstrate less promising outcomes. In contrast, the utilization of AbMIL leads to a notable performance improvement. Moreover, when examining the scattered tissue regions across the WSI using the nested multiple instance method, NMIA, we observe the most significant performance enhancement across all techniques with an AUC of 0.678.

14.7.5 Fusing Image and Clinicopathological Data

By fusing clinicopathological and image data with the aim of complementing each other, as explained in subsection 14.4.3, deep learning models should gain a comprehensive understanding of patient conditions. However, it is imperative to acknowledge the potential limitations and challenges inherent in each form of data. While images offer visual cues that might be ambiguous without clinical context, reports may lack the granularity and specificity present in visual data. Clinicopathological data is often compiled from various sources, including physician notes, laboratory test results, and imaging, potentially introducing noise and variability. Conversely, WSI may also contain inherent variability due to factors such as staining variations or tissue preparation techniques.

In our experiments, unexpectedly, the findings displayed in Table 14.7 indicate that histological features may offer more pertinent information for predicting patient outcomes compared to either clinical data alone or their fusion. This finding raises questions about the traditional reliance on clinical

Method	BCG			Recurrence		
	Best Run	$\mu(\sigma)$	**Montecarlo**	**Best Run**	$\mu(\sigma)$	**Montecarlo**
Maj. Voting	0.500	0.497(0.003)	0.440(0.075)	0.421	0.420(0.001)	0.420(0.001)
Mean	0.427	0.417(0.007)	0.446(0.033)	0.615	**0.583(0.023)**	0.613(0.015)
Max	0.520	**0.499(0.011)**	0.487(0.007)	0.555	0.517(0.047)	0.544(0.008)
AbMIL	0.549	0.434(0.077)	0.548(0.007)	0.592	0.500(0.064)	0.593(0.002)
NMIA	**0.678**	0.486(0.111)	**0.612(0.009)**	**0.721**	0.580(0.079)	**0.722(0.003)**
Clinical	0.501	0.471(0.024)	0.498(0.017)	-	-	-
Clinical + AbMIL	0.502	0.456(0.037)	0.491(0.006)	-	-	-
Clinical + NMIA	0.475	0.450(0.036)	0.454(0.008)	-	-	-

Table 14.7: Test AUC performance for D^{20x}_{UROLP} for BCG and recurrence. The table compares different weakly supervised aggregation techniques. We also show results of integrated clinicopathological reports, histological features, and both in isolation. The results show the mean and standard deviation over 5 runs.

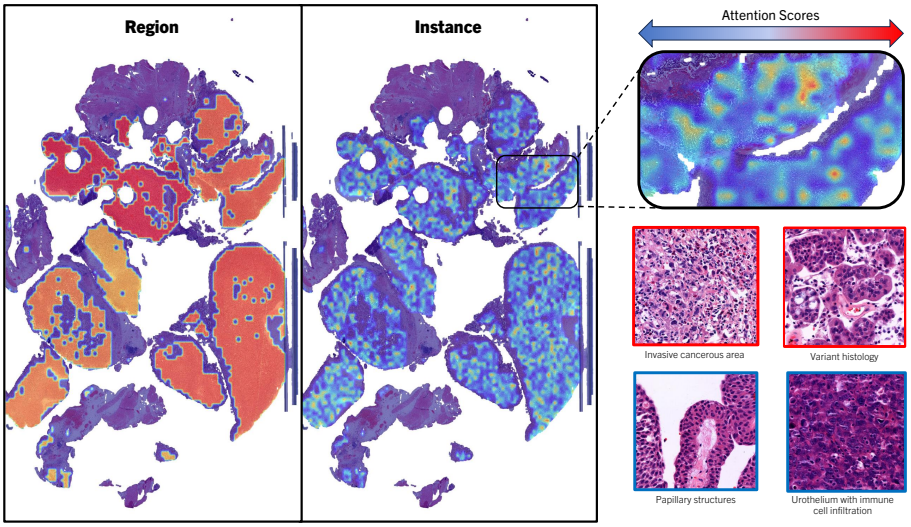


Figure 14.5: Heatmap illustrating attention scores over a BCG-NR WSI from S_{EMC}. The heatmap provides insights into the ROIs where the attention is concentrated within the WSI, facilitating a better understanding of prediction dynamics and highlighting areas of significance for clinical interpretation.

data and highlights the potential of histopathological images as a standalone predictor for improved prognostic accuracy. Moving forward, exploring different fusion methods for integrating clinicopathological and image data could yield insights into the optimal approach for maximizing predictive performance. Moreover, incorporating additional clinical parameters may provide a more comprehensive understanding of the status of the disease.

14.7.6 On the Importance of Manual Annotations

At the inference stage, a fully-independent system cannot be conditioned by manually annotated regions. Even at the training stage, we have observed the challenges in obtaining annotations for all WSI due to the labor-intensive nature of the process. For a prognostic task, the relevance of manual annotations remains uncertain. To explore this aspect, we conducted an experiment where the model was trained on annotated regions ANNO, but tested on automatically segmented regions AUTO, and vice versa. This was compared to a fully automated system for both training and inference. Table 14.8 shows the results, where the ROI-column indicates which ROI definition was used for the AUTO set. For comparison, using ANNO

ROI (Train/Test)	Architecture	URO	LP	UROLP	BORDER	FRONT
ANNO / AUTO	AbMIL	0.489(0.032)	0.446(0.055)	**0.528(0.026)**	0.463(0.054)	0.408(0.025)
	NMIA	0.543(0.022)	0.476(0.054)	0.475(0.032)	0.468(0.054)	0.503(0.027)
AUTO / ANNO	AbMIL	**0.611(0.067)**	0.468(0.144)	0.508(0.090)	**0.600(0.094)**	**0.663(0.098)**
	NMIA	0.486(0.119)	**0.601(0.082)**	0.519(0.054)	0.525(0.054)	0.531(0.073)
AUTO / AUTO	AbMIL	0.492(0.065)	0.452(0.060)	0.434(0.077)	0.534(0.036)	0.484(0.066)
	NMIA	0.535(0.046)	0.513(0.043)	0.486(0.111)	0.512(0.053)	0.428(0.028)

Table 14.8: Comparing AUC performance for distinct training and test datasets, for S_{EMC} 20x magnification. We compare how various models perform when trained on annotations (ANNO) and tested on automatically generated ROIs (AUTO), and vice versa. We also append the test results for models trained and tested with automatically segmented regions. The results show the mean and standard deviation over 5 runs.

for both train and test gives a performance of 0.441(0.031). With one exception, the models perform better after using AUTO-generated ROIs in the training and ANNO for testing, which we interpret as the models benefiting from the larger dataset that is available when we can include non-annotated data.

14.7.7 Recurrence Prediction

Following the predefined settings for ROI, magnification, and feature extraction used in the BCG application, we apply the same configuration for recurrence prediction. The results are presented in Table 14.7. We observe that the utilization of mean aggregation demonstrated the highest average predictive performance. Nevertheless, NMIA exhibited the most substantial performance enhancement among all the techniques studied, with an AUC of 0.721. This underscores the effectiveness of incorporating nested attention mechanisms for the accurate identification of crucial patterns within tissue regions.

14.7.8 Attention-guided Interpretability

The attention scores obtained at the inference stage can be useful for interpretability reasons and can give knowledge on what the model considers relevant for a given prediction. Instances with higher attention scores often correspond to pivotal regions pertinent to the predicted label, aiding in both model validation and reasoning. Attention scores aid pathologists in understanding model's rationale, highlighting areas of interest. This interpretability is valuable for either positive or negative predictions, facilitating validation and refinement of the model's decisions.

An example of attention score heatmap is visualized in Fig. 14.5. Pathologists often use tissue punching to select clinically relevant regions for subsequent analysis in tissue microarrays. While the model consistently allocates its highest attention to punched areas, it is crucial to acknowledge that diagnostically relevant information may extend beyond the punched areas. Nevertheless, the alignment between the model's attention focus and the clinical practice of region selection through punching is an encouraging and promising finding.

14.8 Conclusion

We propose a three-step automated pipeline for the challenging task of prognostic prediction, eliminating the need for manual annotations. This is crucial given the multitude of potential tasks and the limited availability of annotation resources, a trend expected to persist in the future of AI in computational pathology.

The process begins with automated ROI segmentation, where task-specific knowledge can be combined with an automatic tissue classifier to extract relevant ROI. Since prognostic labels are weak in nature, the second step employs contrastive learning to train the feature extractor. Finally, an attention-based nested multiple instance learning classifier, providing predictions and insights into the crucial regions of the image. The pipeline demonstrates exceptional performance on a simpler diagnostic task, achieving a perfect score with an AUC of 1.0 for detecting areas with tumor-infiltrating lymphocytes. Additionally, experiments on synthetic data confirm that the pipeline works perfectly when input distributions are distinct. However, performance drops when they become highly overlapping. In our study, we conducted a thorough pioneering investigation into the use of deep learning and histopathological images for prognostic prediction. We employ response to Bacillus Calmette-Guérin (BCG) treatment in patients with high-risk non-muscle invasive bladder cancer (HR-NMIBC) and recurrence in NMIBC patients as uses cases. The most promising outcomes reveal AUC values of 0.678 for BCG outcome prediction and 0.721 for recurrence prediction. While there is room for improvement, achieving fully automated prognostics based solely on WSI remains a challenging task.

Bibliography

[1] Jo-an Roulson, EW Benbow, and Philip S Hasleton. Discrepancies between clinical and autopsy diagnosis and the value of post mortem histology; a meta-analysis and review. *Histopathology*, 47(6):551–559, 2005.

[2] Jon Griffin and Darren Treanor. Digital pathology in clinical use: where are we now and what is holding us back? *Histopathology*, 70(1):134–145, 2017.

[3] Jonhan Ho, Stefan M Ahlers, Curtis Stratman, Orly Aridor, Liron Pantanowitz, Jeffrey L Fine, John A Kuzmishin, Michael C Montalto, and Anil V Parwani. Can digital pathology result in cost savings? a financial projection for digital pathology implementation at a large integrated health care organization. *Journal of pathology informatics*, 5(1):33, 2014.

[4] Bethany Jill Williams, David Bottoms, and Darren Treanor. Future-proofing pathology: the case for clinical adoption of digital pathology. *Journal of clinical pathology*, 70(12):1010–1018, 2017.

[5] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

[6] J Martin. Meeting pathology demand: Histopathology workforce census. *London: Royal College of Pathologists*, 2018.

[7] Rajesh Singh Laishram. The emerging world of digital pathology. *Journal of Medical Society*, 29:1 – 3, 2015.

[8] Geert Litjens, Francesco Ciompi, and Jeroen van der Laak. A decade of gigascience: the challenges of gigapixel pathology images. *GigaScience*, 11:giac056, 2022.

[9] J Ferlay, M Ervik, F Lam, M Colombet, L Mery, M Piñeros, A Znaor, I Soerjomataram, and F Bray. Global cancer observatory: cancer today. international agency for research on cancer. *Lyon, France*, 2020.

[10] Marko Babjuk, Maximilian Burger, Otakar Capoun, Daniel Cohen, Eva M Compérat, José L Dominguez Escrig, Paolo Gontero, Fredrik Liedberg, Alexandra Masson-Lecomte, A Hugh Mostafid, et al. European association of urology guidelines on non–muscle-invasive bladder cancer (ta, t1, and carcinoma in situ). *European urology*, 81(1):75–94, 2022.

[11] Ashish M Kamat, Richard J Sylvester, Andreas Böhle, Joan Palou, Donald L Lamm, Maurizio Brausi, Mark Soloway, Raj Persad, Roger Buckley, Marc Colombel, et al. Definitions, end points, and clinical trial designs for non–muscle-invasive bladder cancer: recommendations from the international bladder cancer group. *Journal of Clinical Oncology*, 34(16):1935, 2016.

[12] Anastasios Anastasiadis and Theo M de Reijke. Best practice in the treatment of nonmuscle invasive bladder cancer. *Therapeutic advances in urology*, 4(1):13–32, 2012.

[13] Niyati Lobo, Patrick J Hensley, Kelly K Bree, Graciela M Nogueras-Gonzalez, Neema Navai, Colin P Dinney, Richard J Sylvester, and Ashish M Kamat. Updated european association of urology (eau) prognostic factor risk groups overestimate the risk of progression in patients with non–muscle-invasive bladder cancer treated with bacillus calmette-guérin. *European Urology Oncology*, 5(1):84–91, 2022.

[14] Thomas J Fuchs and Joachim M Buhmann. Computational pathology: challenges and promises for tissue analysis. *Computerized Medical Imaging and Graphics*, 35(7-8):515–530, 2011.

[15] David N Louis, Michael Feldman, Alexis B Carter, Anand S Dighe, John D Pfeifer, Lynn Bry, Jonas S Almeida, Joel Saltz, Jonathan Braun, John E Tomaszewski, et al. Computational pathology: a path ahead. *Archives of pathology & laboratory medicine*, 140(1):41–50, 2016.

[16] Esther Abels, Liron Pantanowitz, Famke Aeffner, Mark D Zarella, Jeroen van der Laak, Marilyn M Bui, Venkata NP Vemuri, Anil V Parwani, Jeff Gibbs, Emmanuel Agosto-Arroyo, et al. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the digital pathology association. *The Journal of pathology*, 249(3):286–294, 2019.

[17] Andrew H Song, Guillaume Jaume, Drew FK Williamson, Ming Y Lu, Anurag Vaidya, Tiffany R Miller, and Faisal Mahmood. Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering*, pages 1–20, 2023.

[18] Andrea Duggento, Allegra Conti, Alessandro Mauriello, Maria Guerrisi, and Nicola Toschi. Deep computational pathology in breast cancer. In *Seminars in cancer biology*, volume 72, pages 226–237. Elsevier, 2021.

[19] Andrés Mosquera-Zamudio, Laëtitia Launet, Zahra Tabatabaei, Rafael Parra-Medina, Adrián Colomer, Javier Oliver Moll, Carlos Monteagudo, Emiel Janssen, and Valery Naranjo. Deep learning for skin melanocytic tumors in whole-slide images: A systematic review. *Cancers*, 15(1):42, 2022.

[20] Noémie Rabilloud, Pierre Allaume, Oscar Acosta, Renaud De Crevoisier, Raphael Bourgade, Delphine Loussouarn, Nathalie Rioux-Leclercq, Zine-eddine Khene, Romain Mathieu, Karim Bensalah, et al. Deep learning methodologies applied to digital pathology in prostate cancer: A systematic review. *Diagnostics*, 13(16):2676, 2023.

[21] Ilaria Girolami, Liron Pantanowitz, Stefano Marletta, Meyke Hermsen, Jeroen van der Laak, Enrico Munari, Lucrezia Furian, Fabio Vistoli, Gianluigi Zaza, Massimo Cardillo, et al. Artificial intelligence applications for pre-implantation kidney biopsy pathology practice: a systematic review. *Journal of Nephrology*, 35(7):1801–1808, 2022.

[22] H Mahmood, Muhammad Shaban, BI Indave, AR Santos-Silva, N Rajpoot, and SA Khurram. Use of artificial intelligence in diagnosis of head and neck precancerous and cancerous lesions: a systematic review. *Oral Oncology*, 110:104885, 2020.

[23] Farbod Khoraminia, Saul Fuster, Neel Kanwal, Mitchell Olislagers, Kjersti Engan, Geert JLH van Leenders, Andrew P Stubbs, Farhan Akram, and Tahlita CM Zuiverloon. Artificial intelligence in digital pathology for bladder cancer: Hype or hope? a systematic review. *Cancers*, 15(18):4518, 2023.

[24] Andrew Brodie, Nick Dai, Jeremy Yuen-Chun Teoh, Karel Decaestecker, Prokar Dasgupta, and Nikhil Vasdev. Artificial intelligence in urological oncology: An update and future applications. In *Urologic Oncology: Seminars and Original Investigations*, volume 39, pages 379–399. Elsevier, 2021.

[25] Matteo Ferro, Ugo Giovanni Falagario, Biagio Barone, Martina Maggi, Felice Crocetto, Gian Maria Busetto, Francesco del Giudice, Daniela Terracciano, Giuseppe Lucarelli, Francesco Lasorsa, et al. Artificial intelligence in the advanced diagnosis of bladder cancer-comprehensive literature review and future advancement. *Diagnostics*, 13(13):2308, 2023.

[26] Rune Wetteland, Kjersti Engan, Trygve Eftestøl, Vebjørn Kvikstad, and Emiel AM Janssen. A multiscale approach for whole-slide image segmentation of five tissue classes in urothelial carcinoma slides. *Technology in Cancer Research & Treatment*, 19:1533033820946787, 2020.

[27] Rune Wetteland, Ove Nicolai Dalheim, Vebjørn Kvikstad, Emiel Janssen, and Kjersti Engan. Semi-supervised tissue segmentation of histological images. Colour and visual computing symposium/CEUR Workshop Proceedings, 2020.

[28] Muhammad Khalid Khan Niazi, Enes Yazgan, Thomas E Tavolara, Wencheng Li, Cheryl T Lee, Anil Parwani, and Metin N Gurcan. Semantic segmentation to identify bladder layers from h&e images. *Diagnostic Pathology*, 15(1):1–8, 2020.

[29] Qingyuan Zheng, Rui Yang, Xinmiao Ni, Song Yang, Panpan Jiao, Jiejun Wu, Lin Xiong, Jingsong Wang, Jun Jian, Zhengyu Jiang, et al. Quantitative assessment of tumor-infiltrating lymphocytes using machine learning predicts survival in muscle-invasive bladder cancer. *Journal of Clinical Medicine*, 11(23):7081, 2022.

[30] Uwe Schmidt, Martin Weigert, Coleman Broaddus, and Gene Myers. Cell detection with star-convex polygons. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, pages 265–273. Springer, 2018.

[31] Rune Wetteland, Vebjørn Kvikstad, Trygve Eftestøl, Erlend Tøssebro, Melinda Lillesand, Emiel AM Janssen, and Kjersti Engan. Automatic diagnostic tool for predicting cancer grade in bladder cancer patients using deep learning. *IEEE Access*, 9:115813–115825, 2021.

[32] Wayner Barrios, Behnaz Abdollahi, Manu Goyal, Qingyuan Song, Matthew Suriawinata, Ryland Richards, Bing Ren, Alan Schned, John Seigne, Margaret Karagas, et al. Bladder cancer prognosis using deep neural networks and histopathology images. *Journal of pathology informatics*, 13:100135, 2022.

[33] Zizhao Zhang, Pingjun Chen, Mason McGough, Fuyong Xing, Chunbao Wang, Marilyn Bui, Yuanpu Xie, Manish Sapkota, Lei Cui, Jasreman Dhillon, et al. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nature Machine Intelligence*, 1(5):236–245, 2019.

[34] Peng nien Yin, Kishan Kc, Shishi Wei, Qi Yu, Rui Li, Anne R. Haake, Hiroshi Miyamoto, and Feng Cui. Histopathological distinction of non-invasive and invasive bladder cancers using machine learning approaches. *BMC Medical Informatics and Decision Making*, 20, 2020.

[35] Keishiro Fukumoto, Eiji Kikuchi, Shuji Mikami, Koichiro Ogihara, Kazuhiro Matsumoto, Akira Miyajima, and Mototsugu Oya. Tumor budding, a novel prognostic indicator for predicting stage progression in t1 bladder cancers. *Cancer science*, 107(9):1338–1344, 2016.

[36] Rafael E Jimenez, Edward Gheiler, Peter Oskanian, Rabbi Tiguert, Wael Sakr, David P Wood Jr, J Edson Pontes, and David J Grignon. Grading the invasive component of urothelial carcinoma of the bladder and its relationship with progression-free survival. *The American journal of surgical pathology*, 24(7):980–987, 2000.

[37] Naoto Tokuyama, Akira Saito, Ryu Muraoka, Shuya Matsubara, Takeshi Hashimoto, Naoya Satake, Jun Matsubayashi, Toshitaka Nagao, Aashiq H Mirza, Hans-Peter Graf, et al. Prediction of non-muscle invasive bladder cancer recurrence using machine learning of quantitative nuclear features. *Modern Pathology*, 35(4):533–538, 2022.

[38] Jarle Urdal, Kjersti Engan, Vebjørn Kvikstad, and Emilius AM Janssen. Prognostic prediction of histopathological images by local binary patterns and rusboost. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 2349–2353. IEEE, 2017.

[39] Joaquim Bellmunt, Jaegil Kim, Stephanie A Mullane, Brendan Reardon, Anna Orsola, Eliezer VanAllen, Gad Getz, and David J Kwiatkowski. Genomic predictors of recurrence (r) or progression (p) in high grade t1 (hgt1) non-muscle invasive (nmi) bladder cancer., 2016.

[40] Siteng Chen, Liren Jiang, Xinyi Zheng, Jialiang Shao, Tao Wang, Encheng Zhang, Feng Gao, Xiang Wang, and Junhua Zheng. Clinical use of machine learning-based pathomics signature for diagnosis and survival prediction of bladder cancer. *Cancer science*, 112(7):2905–2914, 2021.

[41] Siteng Chen, Liren Jiang, Encheng Zhang, Shanshan Hu, Tao Wang, Feng Gao, Ning Zhang, Xiang Wang, and Junhua Zheng. A novel nomogram based on machine learning-pathomics signature and neutrophil to lymphocyte ratio for survival prediction of bladder cancer patients. *Frontiers in Oncology*, 11:703033, 2021.

[42] Haoyang Mi, Trinity J Bivalacqua, Max Kates, Roland Seiler, Peter C Black, Aleksander S Popel, and Alexander S Baras. Predictive models of response to neoadjuvant chemotherapy in muscle-invasive bladder cancer using nuclear morphology and tissue architecture. *Cell Reports Medicine*, 2(9), 2021.

[43] Roland Seiler, Hussam Al Deen Ashab, Nicholas Erho, Bas WG van Rhijn, Brian Winters, James Douglas, Kim E Van Kessel, Elisabeth E Fransen van de Putte, Matthew Sommerlad, Natalie Q Wang, et al. Impact of molecular subtypes in muscle-invasive bladder cancer on predicting response and survival after neoadjuvant chemotherapy. *European urology*, 72(4):544–554, 2017.

[44] Andrew Janowczyk and Anant Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 7(1):29, 2016.

[45] Geert Litjens, Clara I Sánchez, Nadya Timofeeva, Meyke Hermsen, Iris Nagtegaal, Iringo Kovacs, Christina Hulsbergen-Van De Kaa, Peter Bult, Bram Van Ginneken, and Jeroen Van Der Laak. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific reports*, 6(1):26286, 2016.

[46] Noorul Wahab, Islam M Miligy, Katherine Dodd, Harvir Sahota, Michael Toss, Wenqi Lu, Mostafa Jahanifar, Mohsin Bilal, Simon Graham, Young Park, et al. Semantic annotation for computational pathology: Multidisciplinary experience and best practice recommendations. *The Journal of Pathology: Clinical Research*, 8(2):116–128, 2022.

[47] Mohamed Amgad, Lamees A Atteya, Hagar Hussein, Kareem Hosny Mohammed, Ehab Hafiz, Maha AT Elsebaie, Ahmed M Alhusseiny, Mohamed Atef AlMoslemany, Abdelmagid M Elmatboly, Philip A Pappalardo, et al. Nucls: A scalable crowdsourcing approach and dataset for nucleus classification and segmentation in breast cancer. *GigaScience*, 11:giac037, 2022.

[48] Ruqayya Awan, Korsuk Sirinukunwattana, David Epstein, Samuel Jefferyes, Uvais Qidwai, Zia Aftab, Imaad Mujeeb, David Snead, and Nasir Rajpoot. Glandular morphometrics for objective grading of colorectal adenocarcinoma histology images. *Scientific reports*, 7(1):16852, 2017.

[49] Jill A Hayden, Danielle A van der Windt, Jennifer L Cartwright, Pierre Côté, and Claire Bombardier. Assessing bias in studies of prognostic factors. *Annals of internal medicine*, 158(4):280–286, 2013.

[50] Robin R Murphy. *Introduction to AI robotics*. MIT press, 2019.

[51] Ira Goldstein and Seymour Papert. Artificial intelligence, language, and the study of knowledge. *Cognitive science*, 1(1):84–123, 1977.

[52] Jaime R Carbonell. Ai in cai: An artificial-intelligence approach to computer-assisted instruction. *IEEE transactions on man-machine systems*, 11(4):190–202, 1970.

[53] Annu Lambora, Kunal Gupta, and Kriti Chopra. Genetic algorithm-a literature review. In *2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon)*, pages 380–384. IEEE, 2019.

[54] Sunday Ayoola Oke. A literature review on artificial intelligence. *International journal of information and management sciences*, 19(4):535–570, 2008.

[55] Mansoureh Maadi, Hadi Akbarzadeh Khorshidi, and Uwe Aickelin. A review on human–ai interaction in machine learning and insights for medical applications. *International journal of environmental research and public health*, 18(4):2121, 2021.

[56] Paul R Daugherty and H James Wilson. *Human+ machine: Reimagining work in the age of AI*. Harvard Business Press, 2018.

[57] Mark O Riedl. Human-centered artificial intelligence and machine learning. *Human behavior and emerging technologies*, 1(1):33–36, 2019.

[58] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[59] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018.

[60] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876, 2021.

[61] Iqbal H Sarker. Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, 2(3):160, 2021.

[62] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[63] Kjersti Engan, Øyvind Meinich-Bache, Sara Brunner, Helge Myklebust, Chunming Rong, Jorge García-Torres, Hege L Ersdal, Anders Johannessen, Hanne Markhus Pike, and Siren Rettedal. Newborn time-improved newborn care based on video and artificial intelligence-study protocol. *BMC Digital Health*, 1(1):1–11, 2023.

[64] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. Deep learning for sensor-based activity recognition: A survey. *Pattern recognition letters*, 119:3–11, 2019.

[65] Weronika Lajewska and Krisztian Balog. Towards filling the gap in conversational search: From passage retrieval to conversational response generation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5326–5330, 2023.

[66] KR Chowdhary. Natural language processing. *Fundamentals of artificial intelligence*, pages 603–649, 2020.

[67] Luca Tomasetti, Kjersti Engan, Mahdieh Khanmohammadi, and Kathinka Dæhli Kurz. Cnn based segmentation of infarcted regions in acute cerebral stroke patients from computed tomography perfusion imaging. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–8, 2020.

[68] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3523–3542, 2021.

[69] Iraia Isasi, Unai Irusta, Elisabete Aramendi, Trygve Eftestøl, Jo Kramer-Johansen, and Lars Wik. Rhythm analysis during cardiopulmonary resuscitation using convolutional neural networks. *Entropy*, 22(6):595, 2020.

[70] Xiaofan Li, Fangwei Dong, Sha Zhang, Weibin Guo, et al. A survey on deep learning techniques in wireless signal recognition. *Wireless Communications and Mobile Computing*, 2019, 2019.

[71] Jiahui Geng, Yongli Mou, Qing Li, Feifei Li, Oya Beyan, Stefan Decker, and Chunming Rong. Improved gradient inversion attacks and defenses in federated learning. *IEEE Transactions on Big Data*, 2023.

[72] Yuandou Wang, Neel Kanwal, Kjersti Engan, Chunming Rong, and Zhiming Zhao. Towards a privacy-preserving distributed cloud service for preprocessing very large medical images. In *2023 IEEE International Conference on Digital Health (ICDH)*, pages 325–327. IEEE, 2023.

[73] Olga Russakovsky, J. Deng, Hao Su, J. Krause, S. Satheesh, S. Ma, Zhiheng Huang, A. Karpathy, A. Khosla, Michael S. Bernstein, A. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015.

[74] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

[75] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021.

[76] Stephanie A. Harmon, Thomas Sanford, G. Thomas Brown, Chris Yang, Sherif Mehralivand, Joseph M. Jacob, Vladimir A. Valera, Joanna H Shih, Piyush Agarwal, Peter L. Choyke, and Baris Turkbey. Multiresolution application of artificial intelligence in digital pathology for prediction of positive lymph nodes from primary tumors in bladder cancer. *JCO clinical cancer informatics*, 4:367–382, 2020.

[77] Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306, 2020.

[78] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[79] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[80] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[81] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

[82] S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109(5):820–838, 2021.

[83] Alana de Santana Correia and E. Colombini. Attention, please! a survey of neural attention models in deep learning. *Artificial Intelligence Review*, 55:6037 – 6124, 2021.

[84] Maximilian Ilse, Jakub M. Tomczak, and M. Welling. Attention-based deep multiple instance learning. In *ICML*, 2018.

[85] QPS Neuropharmacology. Histology services, n.d. Accessed Oct 25, 2023. https://qpsneuro.com/ex-vivo-services/qps-histology-services/.

[86] Giuseppe Musumeci. Past, present and future: overview on histology and histopathology. *J Histol Histopathol*, 1(5):1–3, 2014.

[87] Andrew H Fischer, Kenneth A Jacobson, Jack Rose, and Rolf Zeller. Hematoxylin and eosin staining of tissue and cell sections. *Cold spring harbor protocols*, 2008(5):pdb–prot4986, 2008.

[88] Ada T Feldman and Delia Wolfe. Tissue processing and hematoxylin and eosin staining. *Histopathology: methods and protocols*, pages 31–43, 2014.

[89] Colour Index. The society of dyers and colourists. *Bradford, UK*, 3:4198, 1971.

[90] Rita Maccarone, Stefano Di Marco, and Silvia Bisti. Saffron supplement maintains morphology and function after exposure to damaging light in mammalian retina. *Investigative ophthalmology & visual science*, 49(3):1254–1261, 2008.

[91] Richard Horobin and John Kiernan. *Conn's biological stains: a handbook of dyes, stains and fluorochromes for use in biology and medicine*. Taylor & Francis, 2020.

[92] Roswell Park Compehensive Cancer Center. Bladder cancer staging, n.d. Accessed Oct 25, 2023. https://www.roswellpark.org/cancer/bladder/diagnosis/staging.

[93] Bladder Cancer Academy Network. Bladder cancer types, stages and grades, n.d. Accessed Feb 15, 2024. https://bcan.org/facing-bladder-cancer/bladder-cancer-types-stages-grades/.

[94] DM Wallace, GD Chisholm, and WF Hendry. Tnm classification for urological tumours (uicc)—1974. *British journal of urology*, 47(1):1–12, 1975.

[95] Sven Van Den Bosch and J Alfred Witjes. Long-term cancer-specific survival in patients with high-risk, non–muscle-invasive bladder cancer and tumour progression: a systematic review. *European urology*, 60(3):493–500, 2011.

[96] Richard J Sylvester, Oscar Rodríguez, Virginia Hernández, Diana Turturica, Lenka Bauerová, Harman Max Bruins, Johannes Bründl, Theo H van der Kwast, Antonin Brisuda, José Rubio-Briones, et al. European association of urology (eau) prognostic factor risk groups for non–muscle-invasive bladder cancer (nmibc) incorporating the who 2004/2016 and who 1973 classification systems for grade: an update from the eau nmibc guidelines panel. *European urology*, 79(4):480–488, 2021.

[97] Matthias May, Sabine Brookman-Amissah, Jan Roigas, Arndt Hartmann, Stephan Störkel, Glen Kristiansen, Christian Gilfrich, Roman Borchardt, Bernd Hoschke, Olaf Kaufmann, et al. Prognostic accuracy of individual uropathologists in noninvasive urinary bladder carcinoma: a multicentre study comparing the 1973 and 2004 world health organisation classifications. *European urology*, 57(5):850–858, 2010.

[98] Viktor Soukup, Otakar Čapoun, Daniel Cohen, Virginia Hernandez, Marek Babjuk, Max Burger, Eva Compérat, Paolo Gontero, Thomas Lam, Steven MacLennan, et al. Prognostic performance and reproducibility of the 1973 and 2004/2016 world health organization grading classification systems in non–muscle-invasive bladder cancer: a european association of urology non-muscle invasive bladder cancer guidelines panel systematic review. *European urology*, 72(5):801–813, 2017.

[99] Ok Målfrid Mangrud, Rune Waalen, Einar Gudlaugsson, Ingvild Dalen, Ilker Tasdemir, Emiel AM Janssen, and Jan PA Baak. Reproducibility and prognostic value of who1973 and who2004 grading systems in tat1 urothelial carcinoma of the urinary bladder. *PLoS One*, 9(1):e83192, 2014.

[100] Eva Compérat, Lars Egevad, Antonio Lopez-Beltran, Philippe Camparo, Ferran Algaba, Mahul Amin, Jonathan I Epstein, Hans Hamberg, Christina Hulsbergen-van de Kaa, Glen Kristiansen, et al. An interobserver reproducibility study on invasiveness of bladder cancer using virtual microscopy and heatmaps. *Histopathology*, 63(6):756–766, 2013.

[101] Marko Babjuk, Maximilian Burger, Otakar Capoun, Daniel Cohen, Eva M Compérat, José L Dominguez Escrig, Paolo Gontero, Fredrik Liedberg, Alexandra Masson-Lecomte, A Hugh Mostafid, et al. European association of urology guidelines on non–muscle-invasive bladder cancer (ta, t1, and carcinoma in situ). *European urology*, 2021.

[102] Roberto Contieri, Marco Paciotti, Giovanni Lughezzani, Nicolò M Buffi, Nicola Frego, Pietro Diana, Vittorio Fasulo, Alberto Saita, Paolo Casale, Massimo Lazzeri, et al. Long-term follow-up and factors associated with active surveillance failure for patients with non–muscle-invasive bladder cancer: The bladder cancer italian active surveillance (bias) experience. *European Urology Oncology*, 5(2):251–255, 2022.

[103] Rodolfo Hurle, Luisa Pasini, Massimo Lazzeri, Piergiuseppe Colombo, NicolòMaria Buffi, Giovanni Lughezzani, Paolo Casale, Emanuela Morenghi, Roberto Peschechera, Silvia Zandegiacomo, et al. Active surveillance for low-risk non-muscle-invasive bladder cancer: mid-term results from the bladder cancer italian active surveillance (bias) project. *BJU international*, 118(6):935–939, 2016.

[104] Daniel Ramirez, Amit Gupta, Daniel Canter, Brian Harrow, Ryan W Dobbs, Victor Kucherov, Edward Mueller, Necole Streeper, Matthew A Uhlman, Robert S Svatek, et al. Microscopic haematuria at time of diagnosis is associated with lower disease stage in patients with newly diagnosed bladder cancer. *BJU international*, 117(5):783–786, 2016.

[105] Tony W Trinh, Daniel I Glazer, Cheryl A Sadow, V Anik Sahni, Nina L Geller, and Stuart G Silverman. Bladder cancer diagnosis with ct urography: test characteristics and reasons for false-positive and false-negative results. *Abdominal radiology*, 43:663–671, 2018.

[106] Susan Hilton and Lisa P Jones. Recent advances in imaging cancer of the kidney and urinary tract. *Surgical Oncology Clinics*, 23(4):863–910, 2014.

[107] Valeria Panebianco, Yoshifumi Narumi, Ersan Altun, Bernard H Bochner, Jason A Efstathiou, Shaista Hafeez, Robert Huddart, Steve Kennish, Seth Lerner, Rodolfo Montironi, et al. Multiparametric magnetic resonance imaging for bladder cancer: development of vi-rads (vesical imaging-reporting and data system). *European urology*, 74(3):294–306, 2018.

[108] Madelon NM van der Aa, Ewout W Steyerberg, Chris Bangma, Bas WG van Rhijn, Ellen C Zwarthoff, and Theo H van der Kwast. Cystoscopy revisited as the gold standard for detecting bladder cancer recurrence: diagnostic review bias in the randomized, prospective cefub trial. *The Journal of urology*, 183(1):76–80, 2010.

[109] Maximilian Burger, James W.F. Catto, Guido Dalbagni, Herbert Barton Grossman, Harry W. Herr, Pierre I Karakiewicz, Wassim Kassouf, Lambertus A.L.M. Kiemeney, Carlo la Vecchia, Shahrokh Francois Shariat, and Yair Lotan. Epidemiology and risk factors of urothelial bladder cancer. *European urology*, 63 2:234–41, 2013.

[110] Mostafa Dianatinasab, Anke Wesselius, Amin Salehi-Abargouei, Evan YW Yu, Mohammad Fararouei, Maree Brinkman, Piet van den Brandt, Emily White, Elisabete Weiderpass, Florence Le Calvez-Kelm, et al. Dietary fats and their sources in association with the risk of bladder cancer: a pooled analysis of 11 prospective cohort studies. *International Journal of Cancer*, 151(1):44–55, 2022.

[111] Richard J Sylvester, Adrian PM Van Der Meijden, Willem Oosterlinck, J Alfred Witjes, Christian Bouffioux, Louis Denis, Donald WW Newling, and Karlheinz Kurth. Predicting recurrence and progression in individual patients with stage ta t1 bladder cancer using eortc risk tables: a combined analysis of 2596 patients from seven eortc trials. *European urology*, 49(3):466–477, 2006.

[112] Jesus Fernandez-Gomez, Rosario Madero, Eduardo Solsona, Miguel Unda, Luis Martinez-Piñeiro, Marcelino Gonzalez, Jose Portillo, Antonio Ojea, Carlos Pertusa, Jesus Rodriguez-Molina, et al. Predicting nonmuscle invasive bladder cancer recurrence and progression in patients treated with bacillus calmette-guerin: the cueto scoring model. *The Journal of urology*, 182(5):2195–2203, 2009.

[113] Justin T. Matulay and Ashish M. Kamat. Advances in risk stratification of bladder cancer to guide personalized medicine. *F1000Research*, 7, 2018.

[114] Richard J Sylvester, Willem Oosterlinck, Sten Holmang, Matthew R Sydes, Alison Birtle, Sigurdur Gudjonsson, Cosimo De Nunzio, Kikuo Okamura, Eero Kaasinen, Eduardo Solsona, et al. Systematic review and individual patient data meta-analysis of randomized trials comparing a single immediate instillation of chemotherapy after transurethral resection with transurethral resection alone in patients with stage pta–pt1 urothelial carcinoma of the bladder: which patients benefit from the instillation? *European urology*, 69(2):231–244, 2016.

[115] Stefanie Schmidt, Frank Kunath, Bernadette Coles, Desiree Louise Draeger, Laura-Maria Krabbe, Rick Dersch, Samuel Kilian, Katrin Jensen, Philipp Dahm, and Joerg J Meerpohl. Intravesical bacillus calmette-guérin versus mitomycin c for ta and t1 bladder cancer. *Cochrane Database of Systematic Reviews*, (1), 2020.

[116] MD Shelley, TJ Wilt, J Court, B Coles, H Kynaston, and MD Mason. Intravesical bacillus calmette-guerin is superior to mitomycin c in reducing tumour recurrence in high-risk superficial bladder cancer: a meta-analysis of randomized trials. *BJU international*, 93(4):485–490, 2004.

[117] Per-Uno Malmström, Richard J Sylvester, David E Crawford, Martin Friedrich, Susanne Krege, Erkki Rintala, Eduardo Solsona, Savino M Di Stasi, and J Alfred Witjes. An individual patient data meta-analysis of the long-term outcome of randomised studies comparing intravesical mitomycin c versus bacillus calmette-guérin for non–muscle-invasive bladder cancer. *European urology*, 56(2):247–256, 2009.

[118] Miao Cui and David Zhang. Artificial intelligence and computational pathology. *Laboratory Investigation; a Journal of Technical Methods and Pathology*, pages 1 – 11, 2021.

[119] Lei Cong, Wanbing Feng, Zhigang Yao, Xiaoming Zhou, and Wei Xiao. Deep learning model as a new trend in computer-aided diagnosis of tumor pathology for lung cancer. *Journal of Cancer*, 11(12):3615, 2020.

[120] Rune Wetteland, Kjersti Engan, and Trygve Eftesol. Parameterized extraction of tiles in multilevel gigapixel images. *2021 12th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pages 78–83, 2021.

[121] Shuyan Liu, Junda Ren, Zhineng Chen, Kai Hu, Fen Xiao, Xuanya Li, and Xieping Gao. Effidiag: an efficient framework for breast cancer diagnosis in multi-gigapixel whole slide images. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 663–669. IEEE, 2020.

[122] Sajid Javed, Arif Mahmood, Naoufel Werghi, Ksenija Benes, and Nasir Rajpoot. Multiplex cellular communities in multi-gigapixel colorectal cancer histology images for tissue phenotyping. *IEEE Transactions on Image Processing*, 29:9204–9219, 2020.

[123] Jeroen Van der Laak, Geert Litjens, and Francesco Ciompi. Deep learning in histopathology: the path to the clinic. *Nature medicine*, 27(5):775–784, 2021.

[124] Ankush Patel, Ulysses GJ Balis, Jerome Cheng, Zaibo Li, Giovanni Lujan, David S McClintock, Liron Pantanowitz, and Anil Parwani. Contemporary whole slide imaging devices and their applications within the modern pathology department: A selected hardware review. *Journal of Pathology Informatics*, 12(1):50, 2021.

[125] Maryam Haghighat, Lisa Browning, Korsuk Sirinukunwattana, Stefano Malacrino, Nasullah Khalid Alham, Richard Colling, Ying Cui, Emad Rakha, Freddie C Hamdy, Clare Verrill, et al. Automated quality assessment of large digitised histology cohorts by artificial intelligence. *Scientific Reports*, 12(1):5002, 2022.

[126] Neel Kanwal, Fernando Pérez-Bueno, Arne Schmidt, Kjersti Engan, and Rafael Molina. The devil is in the details: Whole slide image acquisition and processing for artifacts detection, color variation, and data augmentation: A review. *IEEE Access*, 10:58821–58844, 2022.

[127] Massimo Salvi, U Rajendra Acharya, Filippo Molinari, and Kristen M Meiburger. The impact of pre-and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis. *Computers in Biology and Medicine*, 128:104129, 2021.

[128] Alexander I Wright, Catriona M Dunn, Michael Hale, Gordon GA Hutchins, and Darren E Treanor. The effect of quality control on accuracy of digital pathology image analysis. *IEEE Journal of Biomedical and Health Informatics*, 25(2):307–314, 2020.

[129] Neel Kanwal, Saul Fuster, Farbod Khoraminia, Tahlita CM Zuiverloon, Chunming Rong, and Kjersti Engan. Quantifying the effect of color processing on blood and damaged tissue detection in whole slide images. In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, pages 1–5. IEEE, 2022.

[130] Neel Kanwal, Miguel López-Pérez, Umay Kiraz, Tahlita CM Zuiverloon, Rafael Molina, and Kjersti Engan. Are you sure it's an artifact? artifact detection and uncertainty quantification in histological images. *Computerized Medical Imaging and Graphics*, 112:102321, 2024.

[131] Zhongling Wang, Mahdi S Hosseini, Adyn Miles, Konstantinos N Plataniotis, and Zhou Wang. Focuslitenn: High efficiency focus quality assessment for digital pathology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 403–413. Springer, 2020.

[132] Morteza Babaie and Hamid R Tizhoosh. Deep features for tissue-fold detection in histopathology images. In *Digital Pathology: 15th European Congress, ECDP 2019, Warwick, UK, April 10–13, 2019, Proceedings 15*, pages 125–132. Springer, 2019.

[133] Z Swiderska-Chadaj, T Markiewicz, J Gallego, G Bueno, B Grala, and M Lorent. Deep learning for damaged tissue detection and segmentation in ki-67 brain tumor specimens based on the u-net model. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, pages 849–856, 2018.

[134] Gabriele Campanella, Arjun R Rajanna, Lorraine Corsale, Peter J Schüffler, Yukako Yagi, and Thomas J Fuchs. Towards machine learned quality control: A benchmark for sharpness quantification in digital pathology. *Computerized Medical Imaging and Graphics*, 65:142–151, 2018.

[135] Ali RN Avanaki, Kathryn S Espig, Albert Xthona, Christian Lanciault, and Tom RL Kimpe. Automatic image quality assessment for digital pathology. In *Breast Imaging: 13th International Workshop, IWDM 2016, Malmö, Sweden, June 19-22, 2016, Proceedings 13*, pages 431–438. Springer, 2016.

[136] David Ameisen, Christophe Deroulers, Valérie Perrier, Fatiha Bouhidel, Maxime Battistella, Luc Legrès, Anne Janin, Philippe Bertheau, and Jean-Baptiste Yunès. Towards better digital pathology workflows: programming libraries for high-speed sharpness assessment of whole slide images. In *Diagnostic pathology*, volume 9, pages 1–7. BioMed Central, 2014.

[137] Ana Jiménez, Gloria Bueno, Gabriel Cristóbal, Oscar Déniz, David Toomey, and Catherine Conway. Image quality metrics applied to digital pathology. In *Optics, Photonics and Digital Technologies for Imaging Applications IV*, volume 9896, pages 170–187. SPIE, 2016.

[138] David Tellez, Geert Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen Van Der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical image analysis*, 58:101544, 2019.

[139] Mahendra Khened, Avinash Kori, Haran Rajkumar, Ganapathy Krishnamurthi, and Balaji Srinivasan. A generalized deep learning framework for whole-slide image segmentation and analysis. *Scientific reports*, 11(1):11579, 2021.

[140] Marc Macenko, Marc Niethammer, James S Marron, David Borland, John T Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E Thomas. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE international symposium on biomedical imaging: from nano to macro*, pages 1107–1110. IEEE, 2009.

[141] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001.

[142] Ryan A Hoffman, Sonal Kothari, and May D Wang. Comparison of normalization algorithms for cross-batch color segmentation of histopathological images. In *2014 36th annual international conference of the IEEE engineering in medicine and biology society*, pages 194–197. IEEE, 2014.

[143] Abhishek Vahadane, Tingying Peng, Amit Sethi, Shadi Albarqouni, Lichao Wang, Maximilian Baust, Katja Steiger, Anna Melissa Schlitter, Irene Esposito, and Nassir Navab. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE transactions on medical imaging*, 35(8):1962–1971, 2016.

[144] Fernando Pérez-Bueno, Miguel Vega, María A Sales, José Aneiros-Fernández, Valery Naranjo, Rafael Molina, and Aggelos K Katsaggelos. Blind color deconvolution, normalization, and classification of histological images using general super gaussian priors and bayesian inference. *Computer Methods and Programs in Biomedicine*, 211:106453, 2021.

[145] Fernando Pérez-Bueno, Juan G Serra, Miguel Vega, Javier Mateos, Rafael Molina, and Aggelos K Katsaggelos. Bayesian k-svd for h and e blind color deconvolution. applications to stain normalization, data augmentation and cancer classification. *Computerized Medical Imaging and Graphics*, 97:102048, 2022.

[146] Arnout C Ruifrok, Dennis A Johnston, et al. Quantification of histochemical staining by color deconvolution. *Analytical and quantitative cytology and histology*, 23(4):291–299, 2001.

[147] David Morrison, David Harris-Birtill, and Peter D Caie. Generative deep learning in digital pathology workflows. *The American Journal of Pathology*, 191(10):1717–1723, 2021.

[148] Laya Jose, Sidong Liu, Carlo Russo, Annemarie Nadort, and Antonio Di Ieva. Generative adversarial networks in digital pathology and histopathological image processing: A review. *Journal of Pathology Informatics*, 12(1):43, 2021.

[149] Andrew Janowczyk, Ren Zuo, Hannah Gilmore, Michael Feldman, and Anant Madabhushi. Histoqc: an open-source quality control tool for digital pathology slides. *JCO clinical cancer informatics*, 3:1–7, 2019.

[150] Rodrigo Escobar Díaz Guerrero, Lina Carvalho, Thomas Bocklitz, Juergen Popp, and José Luis Oliveira. Software tools and platforms in digital pathology: a review for clinicians and computer scientists. *Journal of Pathology Informatics*, 13:100103, 2022.

[151] Stuart Berg, Dominik Kutra, Thorben Kroeger, Christoph N Straehle, Bernhard X Kausler, Carsten Haubold, Martin Schiegg, Janez Ales, Thorsten Beier, Markus Rudy, et al. Ilastik: interactive machine learning for (bio) image analysis. *Nature methods*, 16(12):1226–1232, 2019.

[152] Michael David Abràmoff, Paulo Jorge Magalhães, and Sunanda J. Ram. Image processing with imagej. 2004.

[153] Anne E Carpenter, Thouis R Jones, Michael R Lamprecht, Colin Clarke, In Han Kang, Ola Friman, David A Guertin, Joo Han Chang, Robert A Lindquist, Jason Moffat, et al. Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome biology*, 7:1–11, 2006.

[154] Peter Bankhead, Maurice B. Loughrey, José A Fernández, Yvonne Dombrowski, Darragh G. McArt, Philip D J Dunne, Stephen McQuaid, Ronan T. Gray, Liam J. Murray, Helen G Coleman, Jacqueline A. James, Manuel Salto-Tellez, and Peter Hamilton. Qupath: Open source software for digital pathology image analysis. *Scientific Reports*, 7, 2017.

[155] Jeff Sevigny, Ping Chiao, Thierry Bussière, Paul H Weinreb, Leslie Williams, Marcel Maier, Robert Dunstan, Stephen Salloway, Tianle Chen, Yan Ling, et al. The antibody aducanumab reduces aβ plaques in alzheimer's disease. *Nature*, 537(7618):50–56, 2016.

[156] Longze Zhang, Martin Chang, Christopher A Beck, Edward M Schwarz, and Brendan F Boyce. Analysis of new bone, cartilage, and fibrosis tissue in healing murine allografts using whole slide imaging and a new automated histomorphometric algorithm. *Bone research*, 4(1):1–9, 2016.

[157] Ann Taber, Emil Christensen, Philippe Lamy, Iver Nordentoft, Frederik Prip, Sia Viborg Lindskrog, Karin Birkenkamp-Demtröder, Trine Line Hauge Okholm, Michael Knudsen, Jakob Skou Pedersen, et al. Molecular correlates of cisplatin-based chemotherapy response in muscle invasive bladder cancer by integrated multi-omics analysis. *Nature communications*, 11(1):4858, 2020.

[158] Bindu Challa, Maryam Tahir, Yan Hu, David Kellough, Giovani Lujan, Shaoli Sun, Anil V Parwani, and Zaibo Li. Artificial intelligence–aided diagnosis of breast cancer lymph node metastasis on histologic slides in a digital workflow. *Modern Pathology*, 36(8):100216, 2023.

[159] S Larry Goldenberg, Guy Nir, and Septimiu E Salcudean. A new era: artificial intelligence and machine learning in prostate cancer. *Nature Reviews Urology*, 16(7):391–403, 2019.

[160] Shidan Wang, Donghan M Yang, Ruichen Rong, Xiaowei Zhan, Junya Fujimoto, Hongyu Liu, John Minna, Ignacio Ivan Wistuba, Yang Xie, and Guanghua Xiao. Artificial intelligence in lung cancer pathology image analysis. *Cancers*, 11(11):1673, 2019.

[161] Marit Lucas, Ilaria Jansen, Ton G. van Leeuwen, Jorg R Oddens, Daniel Martijn de Bruin, and Henk A. Marquering. Deep learning-based recurrence prediction in patients with non-muscle-invasive bladder cancer. *European urology focus*, 2020.

[162] Y. LeCun, L. Bottou, Yoshua Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278–2324, 1998.

[163] Babak Ehteshami Bejnordi, M. Veta, Paul Johannes van Diest, B. van Ginneken, N. Karssemeijer, G. Litjens, J. A. van der Laak, M. Hermsen, Quirine F Manson, and Maschenka C. A. Balkenhol et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318:2199–2210, 2017.

[164] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*, pages 210–218. Springer, 2018.

[165] OpenSeadragon-An Open-Source. Web-based viewer for high-resolution zoomable images, implemented in pure javascript, for desktop and mobile, 2019. Accessed Feb 16, 2024. https://openseadragon.github.io/.

[166] Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*, pages 399–407. Springer, 2017.

[167] Scott Doyle, James Monaco, Michael Feldman, John Tomaszewski, and Anant Madabhushi. An active learning based classification strategy for the minority class problem: application to histopathology annotation. *BMC bioinformatics*, 12:1–14, 2011.

[168] Asim Smailagic, Pedro Costa, Hae Young Noh, Devesh Walawalkar, Kartik Khandelwal, Adrian Galdran, Mostafa Mirshekari, Jonathon Fagert, Susu Xu, Pei Zhang, et al. Medal: Accurate and robust deep active learning for medical image analysis. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, pages 481–488. IEEE, 2018.

[169] Xu Jin, Hong An, Jue Wang, Ke Wen, and Zheng Wu. Reducing the annotation cost of whole slide histology images using active learning. In *2021 3rd International Conference on Image Processing and Machine Vision (IPMV)*, pages 47–52, 2021.

[170] Wenyuan Li, Jiayun Li, Zichen Wang, Jennifer Polson, Anthony E Sisk, Dipti P Sajed, William Speier, and Corey W Arnold. Pathal: An active learning framework for histopathology image analysis. *IEEE Transactions on Medical Imaging*, 41(5):1176–1187, 2021.

[171] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR, 2017.

[172] Brendon Lutnick, Brandon Ginley, Darshana Govind, Sean D McGarry, Peter S LaViolette, Rabi Yacoub, Sanjay Jain, John E Tomaszewski, Kuang-Yu Jen, and Pinaki Sarder. An integrated iterative annotation technique for easing neural network training in medical image analysis. *Nature machine intelligence*, 1(2):112–119, 2019.

[173] Alessandro Tibo, Paolo Frasconi, and Manfred Jaeger. A network architecture for multi-multi-instance learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part I 10*, pages 737–752. Springer, 2017.

[174] Alessandro Tibo, M. Jaeger, and P. Frasconi. Learning and interpreting multi-multi-instance learning networks. *J. Mach. Learn. Res.*, 21:193:1–193:60, 2020.

[175] Ilaria Jansen, Marit Lucas, Judith Bosschieter, Onno J de Boer, Sybren L Meijer, Ton G van Leeuwen, Henk A Marquering, Jakko A Nieuwenhuijzen, Daniel M de Bruin, and C Dilara Savci-Heijink. Automated detection and grading of non–muscle-invasive urothelial cell carcinoma of the bladder. *The American journal of pathology*, 190(7):1483–1490, 2020.

[176] I Tosoni, Urs Wagner, Guido Sauter, Michael Egloff, Hartmut Knönagel, Göran Alund, Fridolin Bannwart, Michael J. Mihatsch, T. Christian Gasser, and Robert Maurer. Clinical significance of interobserver differences in the staging and grading of superficial bladder cancer. *BJU International*, 85, 2000.

[177] Anant Madabhushi and George Lee. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical image analysis*, 33:170–175, 2016.

[178] Amelie Echle, Niklas Rindtorff, Titus Josef Brinker, Tom Luedde, Alexander T. Pearson, and Jakob Nikolas Kather. Deep learning in cancer pathology: a new generation of clinical biomarkers. *British journal of cancer*, 2020.

[179] Soheila Borhani, Reza Borhani, and André Alexander Kajdacsy-Balla. Artificial intelligence: A promising frontier in bladder cancer diagnosis and outcome prediction. *Critical reviews in oncology/hematology*, page 103601, 2022.

[180] Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17, 2015.

[181] Qingyuan Zheng, Rui Yang, Xinmiao Ni, Song Yang, Lin Xiong, Dandan Yan, Lingli Xia, Jingping Yuan, Jingsong Wang, Panpan Jiao, et al. Accurate diagnosis and survival prediction of bladder cancer using deep learning on histological slides. *Cancers*, 14(23):5807, 2022.

[182] Karl-Erik Andersson and Karen D McCloskey. Lamina propria: the functional center of the bladder? *Neurourology and urodynamics*, 33(1):9–16, 2014.

[183] Michael Brooks, Qianxing Mo, Ross Krasnow, Philip Levy Ho, Yu-Cheng Lee, Jing Xiao, Antonina Kurtova, Seth Lerner, Gui Godoy, Weiguo Jian, et al. Positive association of collagen type i with non-muscle invasive bladder cancer progression. *Oncotarget*, 7(50):82609, 2016.

[184] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10, 1997.

[185] Gwénolé Quellec, Guy Cazuguel, Béatrice Cochener, and Mathieu Lamard. Multiple-instance learning for medical image and video analysis. *IEEE Reviews in Biomedical Engineering*, 10:213–234, 2017.

[186] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael A. Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89, 2018.

[187] Gabrielle Ras, Ning Xie, Marcel van Gerven, and Derek Doran. Explainable deep learning: A field guide for the uninitiated. *Journal of Artificial Intelligence Research*, 73:329–397, 2022.

[188] Rusheng Li, Hanhui Liu, Yuesheng Zhu, and Zhiqiang Bai. Arnet: Attention-based refinement network for few-shot semantic segmentation. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2238–2242, 2020.

[189] Changlu Guo, Marton Szemenyei, Yugen Yi, and W. Zhou. Channel attention residual u-net for retinal vessel segmentation. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1185–1189, 2021.

[190] Lian Xu, M. Bennamoun, Farid Boussaïd, S. An, and Ferdous Sohel. An improved approach to weakly supervised semantic segmentation. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1897–1901, 2019.

[191] Sixin Hong, Yuexian Zou, Wenwu Wang, and Meng Cao. Weakly labelled audio tagging via convolutional networks with spatial and channel-wise attention. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 296–300, 2020.

[192] Dawid Rymarczyk, Adriana Borowa, Jacek Tabor, and Bartosz Zieliński. Kernel self-attention for weakly-supervised image classification using deep multiple instance learning. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1720–1729, 2021.

[193] J. Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artif. Intell.*, 201:81–105, 2013.

[194] M. Carbonneau, V. Cheplygina, Eric Granger, and G. Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognit.*, 77:329–353, 2018.

[195] Chi-Long Chen, Chi-Chung Chen, Wei-Hsiang Yu, Szu-Hua Chen, Yu-Chan Chang, T. Hsu, M. Hsiao, Chao-Yuan Yeh, and Cheng yu Chen. An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning. *Nature Communications*, 12, 2021.

[196] Philip Chikontwe, Meejeong Kim, Soo Jeong Nam, Heounjeong Go, and Sang Hyun Park. Multiple instance learning with center embeddings for histopathology classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*, pages 519–528. Springer, 2020.

[197] Chengkun He, Jie Shao, Jiasheng Zhang, and Xiangmin Zhou. Clustering-based multiple instance learning with multi-view feature. *Expert Syst. Appl.*, 162:113027, 2020.

[198] F. Khalvati, Junjie Zhang, A. Wong, and M. Haider. Bag of bags: Nested multi instance classification for prostate cancer detection. *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 146–151, 2016.

[199] M. Lu, Drew F. K. Williamson, Tiffany Y Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data efficient and weakly supervised computational pathology on whole slide images. *Nature biomedical engineering*, 2021.

[200] Jiayun Li, Wenyuan Li, A. Sisk, H. Ye, W. Wallace, W. Speier, and C. Arnold. A multi-resolution model for histopathology image classification and localization with multiple instance learning. *Computers in biology and medicine*, 131:104253, 2021.

[201] Antoine Pirovano, Hippolyte Heuberger, S. Berlemont, Saïd Ladjal, and I. Bloch. Automatic feature selection for improved interpretability on whole slide imaging. *Mach. Learn. Knowl. Extr.*, 3:243–262, 2021.

[202] Yash Sharma, Aman Shrivastava, Lubaina Ehsan, Christopher A Moskaluk, Sana Syed, and Donald Brown. Cluster-to-conquer: A framework for end-to-end multi-instance learning for whole slide image classification. In *Medical Imaging with Deep Learning*, pages 682–698. PMLR, 2021.

[203] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *ArXiv*, abs/1901.06706, 2019.

[204] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pages 265–283, 2016.

[205] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71:209 – 249, 2021.

[206] Samuel Budd, Emma C Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, 71:102062, 2021.

[207] Daisuke Komura and Shumpei Ishikawa. Machine learning methods for histopathological image analysis. *Computational and structural biotechnology journal*, 16:34–42, 2018.

[208] K Stacke, G Eilertsen, J Unger, and C Lundström. A closer look at domain shift for deep learning in histopathology. arxiv. *arXiv preprint arXiv:1909.11575*, 10, 2019.

[209] Burr Settles. Active learning literature survey. *Technical Report*, 2009.

[210] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40, 2021.

[211] Gary S Collins, Paula Dhiman, Constanza L Andaur Navarro, Jie Ma, Lotty Hooft, Johannes B Reitsma, Patricia Logullo, Andrew L Beam, Lily Peng, Ben Van Calster, et al. Protocol for development of a reporting guideline (tripod-ai) and risk of bias tool (probast-ai) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ open*, 11(7):e048008, 2021.

[212] Sandra Morales, Kjersti Engan, and Valery Naranjo. Artificial intelligence in computational pathology–challenges and future directions. *Digital Signal Processing*, 119:103196, 2021.

[213] Yazhou Yang and Marco Loog. Active learning using uncertainty information. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2646–2651. IEEE, 2016.

[214] Ziya Kırkalı, Theresa Y. Chan, Murugesan Manoharan, Ferrán Algaba, Christer Busch, Liang Cheng, Lambertus A. Kiemeney, Martin Kriegmair, Rodolfo Montironi, William M. Murphy, Isabell A. Sesterhenn, Masaaki Tachibana, and Jeff Weider. Bladder cancer: epidemiology, staging and grading, and diagnosis. *Urology*, 66 6 Suppl 1:4–34, 2005.

[215] Donald S. Kaufman, William U. Shipley, and Adam S. Feldman. Bladder cancer. *The Lancet*, 374:239–249, 2009.

[216] Hartwig E. Schwaibold, Sivaprakasam Sivalingam, Florian May, and Rudolf Hartung. The value of a second transurethral resection for t1 bladder cancer. *BJU International*, 97, 2006.

[217] Martin J Magers, Antonio Lopez-Beltran, Rodolfo Montironi, Sean R. Williamson, Hristos Z. Kaimakliotis, and Liang Cheng. Staging of bladder cancer. *Histopathology*, 74:112 – 134, 2019.

[218] Roni M. Cox and Jonathan I. Epstein. Large nested variant of urothelial carcinoma: 23 cases mimicking von brunn nests and inverted growth pattern of noninvasive papillary urothelial carcinoma. *The American Journal of Surgical Pathology*, 35:1337–1342, 2011.

[219] Julio Silva-Rodríguez, Adrián Colomer, María Á. Sales, Rafael Molina, and Valery Naranjo. Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. *Computer methods and programs in biomedicine*, 195:105637, 2020.

[220] Baris Gecer, Selim Aksoy, Ezgi Mercan, Linda G. Shapiro, Donald L. Weaver, and Joann G. Elmore. Detection and classification of cancer in whole slide breast histopathology images using deep convolutional networks. *Pattern recognition*, 84:345–356, 2018.

[221] Ann-Christin Woerl, Markus Eckstein, Josephine Geiger, Daniel-Christoph Wagner, Tamas Daher, Philipp Stenzel, Aurélie Fernandez, Arndt Hartmann, Michael Wand, Wilfried Roth, and Sebastian Foersch. Deep learning predicts molecular subtype of muscle-invasive bladder cancer from conventional histopathological slides. *European urology*, 2020.

[222] Gabriel García, Anna Esteve, Adrián Colomer, David Ramos, and Valery Naranjo. A novel self-learning framework for bladder cancer grading using histopathological images. *Computers in biology and medicine*, 138:104932, 2021.

[223] Rune Wetteland, Kjersti Engan, Trygve Eftestøl, Vebjørn Kvikstad, and Emiel A. M. Janssen. A multiscale approach for whole-slide image segmentation of five tissue classes in urothelial carcinoma slides. *Technology in Cancer Research & Treatment*, 19, 2020.

[224] Samuel J. Galgano, Kristin Kelly Porter, Constantine M Burgan, and Soroush Rais-Bahrami. The role of imaging in bladder cancer diagnosis and staging. *Diagnostics*, 10, 2020.

[225] Sankeerth S. Garapati, Lubomir M. Hadjiiski, Kenny H. Cha, Heang-Ping Chan, Elaine M. Caoili, Richard H. Cohan, Alon Z. Weizer, Ajjai Alva, Chintana Paramagul, Jun Wei, and Chuan Zhou. Urinary bladder cancer staging in ct urography using machine learning. *Medical Physics*, 44:5814–5823, 2017.

[226] Xiaopan Xu, Xi Zhang, Qiang Tian, Huanjun Wang, Long-Biao Cui, Shurong Li, Xing Tang, Baojuan Li, José Dolz, Ismail Ben Ayed, Zhengrong Liang, Jing Yuan, Peng Du, Hongbing Lu, and Yang Liu. Quantitative identification of nonmuscle-invasive and muscle-invasive bladder carcinomas: A multiparametric mri radiomics analysis. *Journal of Magnetic Resonance Imaging*, 49, 2019.

[227] M. Khalid Khan, Thomas E. Tavolara, Vidya Chandrakant Arole, Anil V. Parwani, Cheryl Lee, and Metin Nafi Gürcan. Automated t1 bladder risk stratification based on depth of lamina propria invasion from h and e tissue biopsies: a deep learning approach. In *Medical Imaging*, 2018.

[228] Carlo Patriarca, Rodolfo Hurle, Marco Moschini, Massimo Freschi, Piergiuseppe Colombo, Maurizio Colecchia, L. Ferrari, Giorgio Guazzoni, Andrea Conti, Giario Natale Conti, Roberta Lucianò, Tiziana Magnani, and Renzo Colombo. Usefulness of pt1 substaging in papillary urothelial bladder carcinoma. *Diagnostic Pathology*, 11, 2016.

[229] Martín Abadi, Paul Barham, Jianmin Chen, Z. Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek Gordon Murray, Benoit Steiner, Paul A. Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zhang. Tensorflow: A system for large-scale machine learning. *ArXiv*, abs/1605.08695, 2016.

[230] Laurens van der Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[231] Jeremy Yuen-Chun Teoh, Junjie Huang, Wendy Yuet-Kiu Ko, Veeleah Lok, Peter Choi, Chi-Fai Ng, Shomik Sengupta, Hugh Mostafid, Ashish M Kamat, Peter C Black, et al. Global trends of bladder cancer incidence and mortality, and their associations with tobacco use and gross domestic product per capita. *European urology*, 78(6):893–906, 2020.

[232] JN Eble. World health organization classification of tumours. *Pathology and genetics of tumours of the urinary system and male genital organs*, pages 68–69, 2004.

[233] George J Netto, Mahul B Amin, Daniel M Berney, Eva M Compérat, Anthony J Gill, Arndt Hartmann, Santosh Menon, Maria R Raspollini, Mark A Rubin, John R Srigley, et al. The 2022 world health organization classification of tumors of the urinary system and male genital organs—part b: prostate and urinary tract tumors. *European urology*, 2022.

[234] Anouk E Hentschel, Bas WG van Rhijn, Johannes Bründl, Eva M Compérat, Karin Plass, Oscar Rodríguez, Jose D Subiela Henríquez, Virginia Hernández, Enrique de la Peña, Isabel Alemany, et al. Papillary urothelial neoplasm of low malignant potential (pun-lmp): Still a meaningful histo-pathological grade category for ta, noninvasive bladder tumors in 2019? In *Urologic Oncology: Seminars and Original Investigations*, volume 38, pages 440–448. Elsevier, 2020.

[235] Timothy D Jones and Liang Cheng. Reappraisal of the papillary urothelial neoplasm of low malignant potential (punlmp). *Histopathology*, 77(4):525–535, 2020.

[236] Ayesha S Azam, Islam M Miligy, Peter KU Kimani, Heeba Maqbool, Katherine Hewitt, Nasir M Rajpoot, and David RJ Snead. Diagnostic concordance and discordance in digital pathology: a systematic review and meta-analysis. *Journal of Clinical Pathology*, 2020.

[237] Vebjørn Kvikstad, Ok Målfrid Mangrud, Einar Gudlaugsson, Ingvild Dalen, Hans Espeland, Jan PA Baak, and Emiel AM Janssen. Prognostic value and reproducibility of different microscopic characteristics in the who grading systems for pta and pt1 urinary bladder urothelial carcinomas. *Diagnostic pathology*, 14:1–8, 2019.

[238] Theo van der Kwast, Fredrik Liedberg, Peter C Black, Ashish Kamat, Bas WG van Rhijn, Ferran Algaba, David M Berman, Arndt Hartmann, Antonio Lopez-Beltran, Hemamali Samaratunga, et al. International society of urological pathology expert opinion on grading of urothelial carcinoma. *European Urology Focus*, 8(2):438–446, 2022.

[239] M Alvaro Berbís, David S McClintock, Andrey Bychkov, Jeroen Van der Laak, Liron Pantanowitz, Jochen K Lennerz, Jerome Y Cheng, Brett Delahunt, Lars Egevad, Catarina Eloy, et al. Computational pathology in 2030: a delphi study forecasting the role of ai in pathology within the next decade. *EBioMedicine*, 88, 2023.

[240] Francesco Ciompi, Oscar Geessink, Babak Ehteshami Bejnordi, Gabriel Silva De Souza, Alexi Baidoshvili, Geert Litjens, Bram Van Ginneken, Iris Nagtegaal, and Jeroen Van Der Laak. The importance of stain normalization in colorectal tissue classification with convolutional networks. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 160–163. IEEE, 2017.

[241] Saul Fuster, Farbod Khoraminia, Umay Kiraz, Neel Kanwal, Vebjørn Kvikstad, Trygve Eftestøl, Tahlita CM Zuiverloon, Emiel AM Janssen, and Kjersti Engan. Invasive cancerous area detection in non-muscle invasive bladder cancer whole slide images. In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, pages 1–5. IEEE, 2022.

[242] Angel Cruz-Roa, Hannah Gilmore, Ajay Basavanhally, Michael Feldman, Shridar Ganesan, Natalie Shih, John Tomaszewski, Anant Madabhushi, and Fabio González. High-throughput adaptive sampling for whole-slide histopathology image analysis (hashi) via convolutional neural networks: Application to invasive breast cancer detection. *PloS one*, 13(5):e0196828, 2018.

[243] Sara Reis, Patrycja Gazinska, John H Hipwell, Thomy Mertzanidou, Kalnisha Naidoo, Norman Williams, Sarah Pinder, and David J Hawkes. Automated classification of breast cancer stroma maturity from histological images. *IEEE Transactions on Biomedical Engineering*, 64(10):2344–2352, 2017.

[244] M Murat Dundar, Sunil Badve, Gokhan Bilgin, Vikas Raykar, Rohit Jain, Olcay Sertel, and Metin N Gurcan. Computerized classification of intraductal breast lesions using histopathological images. *IEEE Transactions on Biomedical Engineering*, 58(7):1977–1984, 2011.

[245] Christopher Andreassen, Saul Fuster, Helga Hardardottir, Emiel A.M. Janssen, and Kjersti Engan. Deep learning for predicting metastasis on melanoma wsis. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023.

[246] Zahra Tabatabaei, Adrián Colomer, Javier Oliver Moll, and Valery Naranjo. Toward more transparent and accurate cancer diagnosis with an unsupervised cae approach. *IEEE Access*, 11:143387–143401, 2023.

[247] Ruining Deng, Quan Liu, Can Cui, Tianyuan Yao, Jun Long, Zuhayr Asad, R Michael Womick, Zheyu Zhu, Agnes B Fogo, Shilin Zhao, et al. Omni-seg: A scale-aware dynamic network for renal pathological image segmentation. *IEEE Transactions on Biomedical Engineering*, 2023.

[248] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

[249] Ezgi Mercan, Selim Aksoy, Linda G Shapiro, Donald L Weaver, Tad T Brunyé, and Joann G Elmore. Localization of diagnostically relevant regions of interest in whole slide images: a comparative study. *Journal of digital imaging*, 29:496–506, 2016.

[250] Saul Fuster, Trygve Eftestøl, and Kjersti Engan. Nested multiple instance learning with attention mechanisms. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 220–225. IEEE, 2022.

[251] Panagiota Spyridonos, Panagiotis Petalas, Dimitris Glotsos, D Cavouras, Panagiota Ravazoula, and George Nikiforidis. Comparative evaluation of support vector machines and probabilistic neural networks in superficial bladder cancer classification. *Journal of Computational Methods in Sciences and Engineering*, 6(5-6):283–292, 2006.

[252] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. Mdnet: A semantically and visually interpretable medical image diagnosis network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6428–6436, 2017.

[253] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019.

[254] Saul Fuster, Farbod Khoraminia, Trygve Eftestøl, Tahlita Zuiverloon, and Kjersti Engan. Active learning based domain adaptation for tissue segmentation of histopathological images. In *2023 31st European Signal Processing Conference (EUSIPCO)*, pages 1045–1049. IEEE, 2023.

[255] Vebjørn Kvikstad. Better prognostic markers for nonmuscle invasive papillary urothelial carcinomas. 2022.

[256] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[257] Nabila Abraham and Naimul Mefraz Khan. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*, pages 683–687. IEEE, 2019.

[258] Wicher Bergsma. A bias-correction for cramér's v and tschuprow's t. *Journal of the Korean Statistical Society*, 42(3):323–328, 2013.

[259] Haldun Akoğlu. User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18:91 – 93, 2018.

[260] Florus C de Jong, Teemu D Laajala, Robert F Hoedemaeker, Kimberley R Jordan, Angelique CJ van der Made, Egbert R Boevé, Deric KE van der Schoot, Bart Nieuwkamer, Emiel AM Janssen, Tokameh Mahmoudi, et al. Non–muscle-invasive bladder cancer molecular subtypes predict differential response to intravesical bacillus calmette-guérin. *Science Translational Medicine*, 15(697):eabn4118, 2023.

[261] Xi Wang, Hao Chen, Caixia Gan, Huangjing Lin, Qi Dou, Efstratios Tsougenis, Qitao Huang, Muyan Cai, and Pheng-Ann Heng. Weakly supervised deep learning for whole slide lung cancer image analysis. *IEEE transactions on cybernetics*, 50(9):3950–3962, 2019.

[262] Changjiang Zhou, Yi Jin, Yuzong Chen, Shan Huang, Rengpeng Huang, Yuhong Wang, Youcai Zhao, Yao Chen, Lingchuan Guo, and Jun Liao. Histopathology classification and localization of colorectal cancer using global labels by weakly supervised deep learning. *Computerized Medical Imaging and Graphics*, 88:101861, 2021.

[263] Ching-Wei Wang, Cheng-Chang Chang, Yu-Ching Lee, Yi-Jia Lin, Shih-Chang Lo, Po-Chao Hsu, Yi-An Liou, Chih-Hung Wang, and Tai-Kuang Chao. Weakly supervised deep learning for prediction of treatment effectiveness on ovarian cancer from histopathology images. *Computerized Medical Imaging and Graphics*, 99:102093, 2022.

[264] Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, 67:101813, 2021.

[265] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18802–18812, 2022.

[266] Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65:101789, 2020.

[267] Van-Linh Le, Audrey Michot, Amandine Crombé, Carine Ngo, Charles Honoré, Jean-Michel Coindre, Olivier Saut, and Francois Le-Loarer. A deep attention-multiple instance learning framework to predict survival of soft-tissue sarcoma from whole slide images. In *MICCAI Workshop on Cancer Prevention through Early Detection*, pages 3–16. Springer, 2023.

[268] Pei Liu, Luping Ji, Feng Ye, and Bo Fu. Advmil: Adversarial multiple instance learning for the survival analysis on whole-slide images. *Medical Image Analysis*, 91:103020, 2024.

[269] Pei Liu, Bo Fu, Feng Ye, Rui Yang, and Luping Ji. Dsca: A dual-stream network with cross-attention on whole-slide image pyramids for cancer prognosis. *Expert Systems with Applications*, 227:120280, 2023.

[270] Lucy Godson, Navid Alemi, Jérémie Nsengimana, Graham P Cook, Emily L Clarke, Darren Treanor, D Timothy Bishop, Julia Newton-Bishop, Ali Gooya, and Derek Magee. Immune subtyping of melanoma whole slide images using multiple instance learning. *Medical Image Analysis*, 93:103097, 2024.

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

[272] Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, 2022.

[273] Parsa Ashrafi Fashi, Sobhan Hemati, Morteza Babaie, Ricardo Gonzalez, and HR Tizhoosh. A self-supervised contrastive learning approach for whole slide image representation in digital pathology. *Journal of Pathology Informatics*, 13:100133, 2022.

[274] Jing Ke, Yiqing Shen, Xiaoyao Liang, and Dinggang Shen. Contrastive learning based stain normalization across multiple tumor in histopathology. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, pages 571–580. Springer, 2021.

[275] José Carlos Gutiérrez Pérez, Daniel Otero Baguer, and Peter Maass. Staincut: Stain normalization with contrastive learning. *Journal of Imaging*, 8(7), 2022.

[276] Jun Li, Yushan Zheng, Kun Wu, Jun Shi, Fengying Xie, and Zhiguo Jiang. Lesion-aware contrastive representation learning for histopathology whole slide images analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 273–282. Springer, 2022.

[277] Chao Tu, Yu Zhang, and Zhenyuan Ning. Dual-curriculum contrastive multi-instance learning for cancer prognosis analysis with whole slide images. *Advances in Neural Information Processing Systems*, 35:29484–29497, 2022.

[278] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.