Automated Grading of Bladder Cancer using Deep Learning

by

Rune Wetteland

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
2021

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

ISBN 978-82-8439-056-7
ISSN 1890-1387

PhD Thesis UiS no. 624

Preface

This thesis is submitted as partial fulfillment 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 collaboration with the Department of Pathology, Stavanger University Hospital.

This thesis comprises a collection of five peer-reviewed and published scientific papers. For increased readability, the papers have been reformatted for alignment with the format of the thesis and are included as chapters.

Rune Wetteland, November 2021

Abstract

Urothelial carcinoma is the most common type of bladder cancer and is among the cancer types with the highest recurrence rate and lifetime treatment cost per patient. Diagnosed patients are stratified into risk groups, mainly based on the histological grade and stage. However, it is well known that correct grading of bladder cancer suffers from intra- and interobserver variability and inconsistent reproducibility between pathologists, potentially leading to under- or overtreatment of the patients. The economic burden, unnecessary patient suffering, and additional load on the health care system illustrate the importance of developing new tools to aid pathologists.

With the introduction of digital pathology, large amounts of data have been made available in the form of digital histological whole-slide images (WSI). However, despite the massive amount of data, annotations for the given data are lacking. Another potential problem is that the tissue samples of urothelial carcinoma contain a mixture of damaged tissue, blood, stroma, muscle, and urothelium, where it is mainly the urothelium tissue that is diagnostically relevant for grading.

A method for tissue segmentation is investigated, where the aim is to segment WSIs into the six tissue classes: urothelium, stroma, muscle, damaged tissue, blood, and background. Several methods based on convolutional neural networks (CNN) for tile-wise classification are proposed. Both single-scale and multiscale models were explored to see if including more magnification levels would improve the performance. Different techniques, such as unsupervised learning, semi-supervised learning, and domain adaptation techniques, are explored to mitigate the challenge of missing large quantities of annotated data.

It is necessary to extract tiles from the WSI since it is intractable to process the entire WSI at full resolution at once. We have proposed a method to parameterize and automate the task of extracting tiles from different scales with a region of interest (ROI) defined at one of the scales. The method is reproducible and easy to describe by reporting the parameters.

A pipeline for automated diagnostic grading is proposed, called $\text{TRI}_{\text{grade}}$. First, the tissue segmentation method is utilized to find the diagnostically

relevant urothelium tissue. Then, the parameterized tile extraction method is used to extract tiles from the urothelium regions at three magnification levels from 300 WSIs. The extracted tiles form the training, validation, and test data used to train and test a diagnostic model. The final system outputs a segmented tissue image showing all the tissue regions in the WSI, a WHO grade heatmap indicating low- and high-grade carcinoma regions, and finally, a slide-level WHO grade prediction. The proposed $\text{TRI}_{\text{grade}}$ pipeline correctly graded 45 of 50 WSIs, achieving an accuracy of 90%.

Acknowledgements

First and foremost, I am profoundly grateful to my research supervisor, Professor Kjersti Engan, for all the guidance and unparalleled support throughout this Ph.D. project. The dedication you have for your work is awe-inspiring. No matter how busy, you always have time for me. I couldn't have asked for a better supervisor. I want to extend my deepest gratitude to my co-supervisor, Professor Trygve Eftestøl. Thank you for your insightful and invaluable feedback and for always supporting me. Your comments and suggestions, linked with your keen eye for detail, were of great value to the project.

My sincere thanks to co-supervisor Dr. Emiel Janssen and pathologists Vebjørn Kvikstad at the Stavanger University Hospital. This research would have been impossible without the aid and support from the two of you. It has always been a great joy listening to the knowledge and experience of both of you. My sincere thank you also goes to all my co-authors for your help and contributions. To all my supervisors and co-authors, it has truly been motivating to be surrounded by such a knowledgeable team of people. You have all pushed me to reach a higher level.

I would also like to show gratitude to my leader and head of the department, Dr. Tom Ryen, for your inspiring leadership. Thank you for the opportunities you gave me as the University's representative at NORA's Education Council and for offering me the position as moderator of the GPUs at the University. Thanks also to the Unix administrator Theodor Ivesdal. I learned so much from you and from moderating the GPUs on the Unix network, a work I very much appreciated. Thank you to my fellow Ph.D. candidates and colleagues at the University. And finally, a special thanks to my family and friends for all your support.

Rune Wetteland, November 2021

List of publications

The main part of this dissertation is made up of the following published scientific papers:

- **Paper 1**

  Multiclass Tissue Classification of Whole-Slide Histological Images using Convolutional Neural Networks
  Rune Wetteland, Kjersti Engan, Trygve Eftestøl, Vebjørn Kvikstad, Emiel A. M. Janssen
  Published by SciTePress in the Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods (ICPRAM), 2019.

- **Paper 2**

  A Multiscale Approach for Whole-Slide Image Segmentation of Five Tissue Classes in Urothelial Carcinoma Slides
  Rune Wetteland, Kjersti Engan, Trygve Eftestøl, Vebjørn Kvikstad, Emiel A. M. Janssen
  Published in the Journal of Technology in Cancer Research and Treatment (TCRT), 2020.

- **Paper 3**

  Semi-Supervised Tissue Segmentation of Histological Images
  Ove Nicolai Dalheim, **Rune Wetteland**, Vebjørn Kvikstad, Emiel A. M. Janssen, Kjersti Engan
  Published in the Proceedings of the 10th Colour and Visual Computing Symposium (CVCS), 2020.

- **Paper 4**

  Parameterized Extraction of Tiles in Multilevel Gigapixel Images

Rune Wetteland, Kjersti Engan, Trygve Eftestøl
Published by IEEE in the Proceedings of the 12th International Symposium on Image and Signal Processing and Analysis (ISPA), 2021.

- **Paper 5**

Automatic Diagnostic Tool for Predicting Cancer Grade in Bladder Cancer Patients Using Deep Learning

Rune Wetteland, Vebjørn Kvikstad, Trygve Eftestøl, Erlend Tøssebro, Melinda Lillesand, Emiel A. M. Janssen, Kjersti Engan
Published in the Journal IEEE Access, 2021.

Glossary

AE	Autoencoder
AI	Artificial intelligence
BGR	Blue Green Red (color model)
BMDLab	Biomedical data analysis laboratory
CAD	Computer-aided diagnostic
CNN	Convolutional neural network
CV	Cross-validation
CZI	File format produced by Zeiss slide scanners
DICOM	Digital imaging and communications in medicine
FC	Fully-connected
FCNN	Fully-connected neural network
FDA	U.S. Food and drug administration
FN	False negative
FP	False positive
GAP	Global average pooling
GPU	Graphics processing unit
H&E	Hematoxylin eosin
HCA	Hierarchical cluster analysis
HES	Hematoxylin eosin saffron
JPEG	Joint Photographic Experts Group
MIBC	Muscle-invasive bladder cancer
ML	Machine learning
NDPI	File format produced by Hamamatsu slide scanners
NMIBC	Non-muscle invasive bladder cancer
NN	Neural network

PCA	Principal component analysis
PNG	Portable Network Graphics
RBM	Restricted boltzmann machine
ReLU	Rectified linear unit
RGB	Red Green Blue (Color model)
RNN	Recurrent neural network
ROI	Region-of-interest
SCN	Leica file format
SCN400	Leica's slide scanner
SGD	Stochastic gradient descent
SL	Supervised learning
SSL	Semi-supervised learning
SVM	Support vector machine
TIFF	Tagged Image File Format
TMB	Tumor mutational burden
TN	True negative
TNM	Cancer staging system
TP	True positive
TURBT	Transurethral resection of bladder tumor
VIPS	VASARI image processing system
WHO	World health orginazation
WHO04	WHO 2004 guidelines
WHO16	WHO 2016 guidelines
WHO73	WHO 1973 guidelines
WSI	Whole-slide image
XAI	Explainable artificial intelligence
XML	Extensible markup language

Contents

Chapter 1

Introduction

This chapter will present the background and motivation of the work, followed by an overview of research challenges and opportunities. Next, the thesis objectives and related previous work will be presented. Lastly, the main contributions and outline of the thesis are given.

1.1 Background and motivation

Bladder cancer

There has been a significant increase in both new incidents of urinary bladder cancer and mortalities over the past decades. Globally, there were 573 278 new bladder cancer incidents in 2020 and 212 536 deaths from the disease, making it the 10^{th} most common cancer disease for both sexes combined [119]. Men are overrepresented in this statistics, with approximately 77% and 75% of the incidents and mortalities occurring in men, respectively [119]. This makes bladder cancer the 6th most common cancer disease among men. In addition, bladder cancer is known as one of the most recurring cancer types. Of all patients diagnosed with bladder cancer, 50% to 70% will experience one or more recurrences, and 10% to 30% will have disease progression to a higher stage [81].

There are several kinds of bladder cancer, such as squamous cell carcinoma and adenocarcinoma; however, urothelial carcinoma is by far the most common type, with as much as 90% of the incidents in some regions [37]. Because of the implication of the disease, it requires a very intensive treatment and follow-up plan. Consequently, bladder cancer is one of the cancer types with the highest lifetime treatment cost per patient [12, 111].

In histopathological diagnostics, an expert pathologist will determine the grade and stage of the tumor and describe it according to the latest

WHO16 classification system [7]. Evaluation of the tumor is usually performed manually through a microscope, a time-consuming, and challenging process. The grade and stage of the tumor are used to stratify patients into risk groups, which dictates a suitable treatment and follow-up plan for each patient. However, it is well known that correct grading of bladder cancer suffers from intra- and interobserver variability and inconsistent reproducibility between pathologists [65, 82], potentially leading to under-or overtreatment of the patients.

Among the main challenges for bladder cancer diagnosis is correctly identifying patients at higher risk of recurrence or those facing overtreatment. A patient facing undertreatment may experience recurrence and go a prolonged period before proper treatment, risking both disease progression and the tumor spreading into nearby tissue. In contrast, a patient experiencing overtreatment will undergo unnecessary suffering inflicted by the more vigorous treatment program. Consequently, this will also lead to an additional cost and load on the health care system. The economic burden, unnecessary patient suffering, and additional load on the health care system illustrate the importance of developing new tools to aid pathologists.

As seen in Figure 1.1, there has been a considerable increase in both new incidents and mortalities by bladder cancer in the past two decades. The majority of the numbers belong to men, but an increase is seen in both sexes. To make matters worse, there is a lack of pathologists combined with the ever-growing number of patients. This shortcoming of pathologists could potentially result in less time per patient.

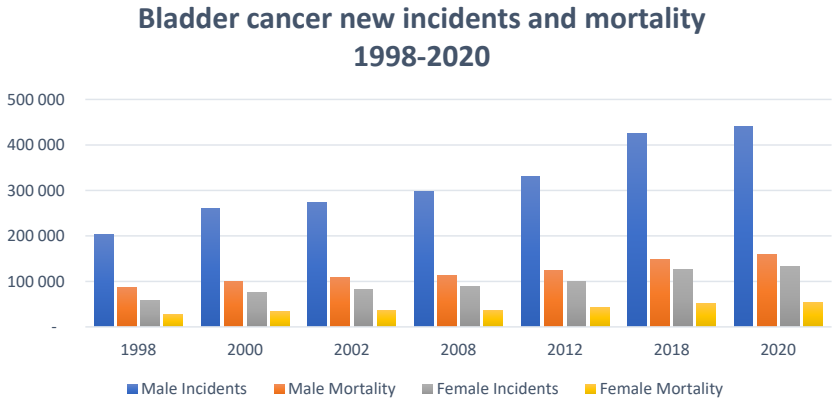


Figure 1.1: Global incidents and mortalities for bladder cancer, both sexes, in the period 1998 to 2020. Data collected from the GLOBOCAN reports [15, 39, 40, 95, 96, 97, 119].

A pathologist's workload is not limited to the grading of urothelium carcinoma, but they have a wide range of other tasks at hand, e.g., diagnosis of other cancer diseases and forensic pathology. They have a critical role, and numerous other disciplines rely on their work. If the workload in one of these tasks increases, such as the increasing number of bladder cancer patients, it may simultaneously affect the time available for other vital tasks.

Märkl et al. [83] recently did a study on the number of German physicians in pathology and compared it with European countries, USA, and Canada. The results indicate a shortage of pathologists in Germany, and a survey revealed an increase in workload for university pathologists over the past decade. The study was primarily focused on the German situation, but the authors state that "*the key findings of this study have implications for many if not all European countries, northern America and probably at least some countries in Asia.*" As a possible solution, the authors hint that automation and new digital technology "*could offer enormous potential for relieving pathologists in their daily work*" [83].

Digital pathology

For over a century, pathologists around the world have been examining tissue specimens through a microscope. However, with the Digital Revolution, otherwise known a the Third Industrial Revolution, novel technologies emerged and shifted the analog technologies to new digital technologies. Within pathology, commercial slide scanners allowed the glass slides to be stored on a computer as a digital image and introduced a new field called digital pathology. This process, referred to as whole-slide imaging (WSI), uses a digital microscope and scans the glass slides at very high resolution.

There are many benefits of utilizing digital versions of the glass sides. Sharing slides between institutes across the world becomes both easier and faster, allowing new collaborations. It opens the possibility for remote access, more accessible storage, and cloud computing. Furthermore, and most importantly, it opens the opportunity for computational pathology, enabling new tools to analyze and interpret the specimen. The significance of this is improved workflow, automation of tedious tasks such as cell counting, potentially improved diagnostic accuracy, and better clinical outcome for the patients [10, 16, 51, 79, 90].

Digital pathology and access to digital slides open new possibilities and is gaining research interest. Morales et al. [86] analyzed the amount of

research related to computational pathology between 2011 and 2020. In PubMed, 4 983 research papers were published, and for Google Scholar, the number was 96 830, with the majority of these papers being published in the last five years.

Lack of annotations

With the introduction of commercial slide scanners and digital pathology, large amounts of data have been made available in the form of histological whole-slide images. However, despite the massive amount of data, annotations for the given data are lacking. That is, a region in a WSI of a known class, e.g., a region of interest (ROI) indicating, for instance, low-grade or muscle tissue. These annotations are important for several reasons; first, they are required to train models using supervised or semi-supervised learning methods. Secondly, it helps reduce the massive amount of data provided by the WSIs to the ROI for the given task, excluding the unwanted areas outside the ROI. Lastly, to evaluate a method, the trained model must be tested on data with known attributes. This means the tested WSIs or ROIs need to be verified by an expert pathologist prior to the test to compare the model's performance against them.

For traditional datasets, like ImageNet, examples of the known classes are cat, dog, car, airplane, and similar. Annotations for these kinds of images are easy to offer, as anybody can provide them. The challenge lies in gathering millions of labels, and methods such as crowdsourcing are used to achieve this. For medical images, however, the main challenge lies in the difficulties in providing the correct label for a given area. This task requires expert input by pathologists and thus is a costly job. Pathologists are in shortage and a limited resource; hence, a large manually annotated dataset is impossible to achieve.

Some possible solutions to this problem can be the use of weak labels, unsupervised learning strategies, or domain adaptation techniques. A weak label is a label that is inaccurate in the description. It may cover a larger area, where parts of this area have a different class than the class label itself. An example of this can be a WSI where a pathologist has graded the entire slide as a specific grade, even though some areas in the image may consist of a different grade. By sampling tiles from this image, one popular strategy is to inherit the slide-level grade to all tiles within the image. Each tile is then weakly defined, as some of the tiles may represent one grade and the associated label to another grade. Unsupervised learning

techniques aim at training models on data without the use of annotations. Domain adaptation techniques use models pre-trained on images from one domain and then fine-tunes it on a new domain. This process reduced the necessary amount of annotated data.

Artificial intelligence

Artificial intelligence and machine learning methods emerged in the 1950s. However, the lack of large annotated datasets and advanced hardware capable of handling the computational complexity of the methods slowed down the development. The field has experienced several AI winters, halting the economic support for continued research.

Following the development of the Digital Revolution, improved computer hardware has allowed machine learning approaches more available for people. In addition, dedicated researchers have curated several large open-source datasets, such as MNIST, ImageNet, CIFAR-10, which helped other researchers focus on developing the machine learning algorithms. Furthermore, the releases of deep learning frameworks, like Tensorflow, Keras, and PyTorch, have made machine learning methods more accessible for researchers. As a result of these combined efforts, artificial intelligence has seen rapid growth in recent years; research interest and commercial products have skyrocketed, with little signs of slowing down.

The interest in AI is not without reason. Time and time again, state-of-the-art results have been set by a machine learning method [66]. For example, the ImageNet contest in 2012 marked the first time a machine learning algorithm contributed to the competition. The model was AlexNet, and it won by a good margin, marking the abandonment of feature engineering for the benefit of machine learning within computer vision tasks [44]. The winner of the 2015 edition of the same competition marked the first model surpassing human-level performance [53]. Similar stories are also seen within the medical field, where deep learning models outperform human expert's performance [17].

Following the general success of machine learning, there is an interest in utilizing such tools to assist pathologists in their work. This is also stated by the authors Morales et al. [86] in their recent overview, where they state that "*the combination of digital histopathology imaging and AI therefore presents a significant opportunity for the support of the pathologists' tasks and opens up a whole new world of computational analysis.*" A successful

computer-aided diagnosis (CAD) system could potentially help improve on the low reproducibility, decrease the variability in interpretations, reduce the increasing workload, and improve the workflow and patient outcome. It has also been shown by Want et al. [133] that errors made by an algorithm and pathologist are different and that the best result was achieved by combining the two.

1.2 Research challenges and opportunities

With the introduction of digital pathology, some computer-aided tools to assist pathologists have been introduced for other diseases. However, such tools are currently not in use for the assessment of urothelial carcinoma slides, which are mainly examined manually through a microscope. This is a time-consuming process, and reproducibility among pathologists is in some cases low, for example, within the prognostic classification of urinary bladder cancer. New tools to aid pathologists in their work are therefore desired. Successful implementation of such tools can improve the workflow, raise the accuracy, and increase the quality of the treatment, thus greatly benefiting the patients suffering from the disease.

The large WSI of urothelium carcinoma contains areas of different tissue types and damaged and burnt tissue areas. Therefore, automatic extraction of diagnostic relevant tissue would be an important step towards automatic grading and prognosis prediction. Prior to this work, no methods were reported for tissue-type segmentation of whole-slide images of urothelial carcinoma.

The large size and multiscale nature of the WSI make it necessary to patch up the images before processing or to extract smaller tiles, possibly over different resolutions. However, to the best of the author's knowledge, a technical description or source code for extracting tiles in multilevel gigapixel images, for example, based on coordinates or masks defining regions of interest, does not exist. This makes reproducibility low if patching and tiling methods can not be described well and parameterized. Therefore, a sound, efficient, parameterized, and automated method for extracting tiles would be useful as a data curation or preprocessing step that can be accurately reported for reproducibility.

As will be discussed in Section 1.4 Previous work, some work on grading urothelial carcinoma slides exist. However, this topic is far from fully explored, and more work is needed both on automatic grading systems, as well as staging and prognosis prediction.

1.3 Thesis objectives

The main goal of this study is to develop an automatic diagnostic system for grading bladder cancer, type urothelium carcinoma. We have access to a dataset of WSIs, with associated slide-level diagnostic labels from Stavanger University Hospital. However, not all parts of the WSI are diagnostically relevant; hence, a tissue segmentation algorithm is necessary to find and extract the diagnostic relevant areas of the WSIs. Furthermore, as detailed region-based annotating in WSIs requires an expert's opinion, we only have access to a small number of annotations for the different tissue classes in the WSIs.

The thesis objectives are divided into one main objective and four sub-objectives as follows:

O_1: Create a system for automated grading of urothelial carcinoma slides.

- SO_1: Create an automated system for distinguishing between the different tissue types present in histological whole-slide images of urothelial carcinoma.
- SO_2: Explore different approaches for unsupervised and semi-supervised learning techniques to deal with the lack of detailed region-based annotation data.
- SO_3: Investigate the use of multiscale models in WSI processing by utilizing several magnification scales.
- SO_4: Create a reproducible system that automatically extracts tiles from multilevel whole-slide images.

1.4 Previous work

This section will look at some of the related works in the field of tissue segmentation and automatic diagnostic methods for bladder cancer.

There exist some related work for multiclass tissue classification on other cancer types [5, 62, 71, 131, 134]. However, to the author's knowledge, there was no published research on multiclass tissue segmentation of urothelial carcinoma WSIs prior to the work presented in this thesis.

There is, however, some work on two-class segmentation of bladder cancer images. These methods aim to classify tiles from the images into one of

two classes, often tumor vs. non-tumor, cancer vs. non-cancer, and similar. For example, in the work of Xu et al. [142], a method for predicting low or high tumor mutational burden (TMB) in bladder cancer patients was investigated. As a preprocessing step, a tile-wise tumor vs. non-tumor classifier was used to segment out the tumor regions from the surrounding tissue. An support-vector machine (SVM) classifier was then used to predict the patient's TMB state using extracted histological image features from the tumor regions. A similar approach was used by Zhang et al. [151], where a U-net-like network was used to predict each pixel into tumor or non-tumor as a preprocessing step before using another neural network for predicting the slide level diagnosis.

The majority of the research on cancer diagnostic follows a two-stage approach. First, a detection algorithm is used to find the diagnostic relevant areas in the images, followed by a classification of this area. This has many advantages, such as reducing the area needed to be processed in the second stage and also removing unwanted tissue classes. This is a quite common methodology, and several researchers have come up with a variety of approaches [19, 59, 78, 115, 151].

In Jansen et al. [59], they utilized two individual single-scale neural networks to detect and grade 328 cases of bladder cancer collected from 232 patients. A U-net-based segmentation network was trained to detect and segment the urothelium tissue, used as input to a second network trained to grade the urothelium tissue according to the WHO04 grading system. The classification network assessed the WHO04 grading on slide-level, using the majority vote of all classified tiles. The predictions were compared with the grading of three experienced pathologists. According to the consensus reading, the classification model achieved an accuracy score of 74%. The included whole-slide images were all exported at 20x magnification (0.5 µm per pixel).

To mimic the work of pathologists, some work utilizes multiscale methods to incorporate both details and context from a broader field of view in the models. This is done by using multiple magnification scales, or by using tiles from the same scale but with varying sizes to accommodate the larger field of view. Reported works by Sirinukunwattana et al. [114], Vu et al. [131], and Hashimoto et al. [52] supports the claim that multiscale models have the potential to improve the classification performance. In Hashimoto et al. [52], the authors also confirm that class-specific features exist at different magnification scales.

Both before and during the work of this thesis, there has been some work on bladder cancer image analysis at our research group at the University of Stavanger in collaboration with Stavanger University Hosptial. The majority of these are master thesis, and includes work on tissue segmentation, prediction of recurrence, detection of cells and immune cells, and assessment of immune cells [31, 80, 120, 126, 127, 128, 136].

The greater part of research on cancer diagnostic are devoted to other cancer types, such as breast, lung, prostate, brain, and skin cancer [88]. This is also the case for AI-based medical technologies approved by the U.S. Food and Drug Administration (FDA), which mostly are in the fields of radiology, cardiology, and Internal Medicine/General Practice [9]. A similar trend can be seen in commercial companies. E.g., ContextVision, one of the leading companies in the field of medical image processing uses artificial intelligence for cancer diagnosis. Within digital pathology, they have products for prostate, lung, and colon. But still, a lot of effort is also aimed towards histological images [19, 25, 45, 57, 117].

One of the main goals of research on bladder cancer is to create new tools to aid pathologists, and a significant amount of work in this thesis is aimed to create helpful and intuitive visualization, which can be used in a clinical setting. Some work presents a few selected close-up areas of segmentation [45], and some work presents segmentation of full WSIs [52, 117, 151]. However, there is no reported work on visualizing all tissue classes from bladder cancer.

1.5 Main contributions

The main contribution of this thesis is an end-to-end diagnostic pipeline for urothelium carcinoma. The pipeline consists of several methods, which are described and presented in five scientific papers. The first three papers are dedicated to the topic of tissue segmentation, Paper 4 covers parameterized and reproducible tile extraction in multilevel gigapixel images, and finally, Paper 5 is about cancer diagnosis, specifically grading of urothelium carcinoma based on WSI input without any manual ROI markings. An overview of the proposed pipeline is depicted in Figure 1.2, indicating at which step each paper contributes. In addition, an overview of all the papers and how they are connected is shown in Figure 1.3.

In Paper 1, an autoencoder model was used to utilize a large dataset of unlabeled data and then fine-tuned on a smaller dataset with annotations.

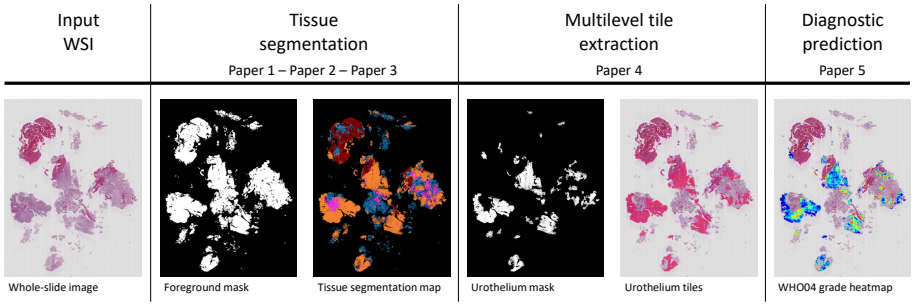


Figure 1.2: An overview of the proposed pipeline. **Input WSI)** A whole-slide image is used as input. **Tissue segmentation)** A foreground mask is used as a reference to extract tiles. Then, a tissue segmentation model is used to identify each tissue class. **Multilevel tile extraction)** The urothelium regions are used to create a urothelium mask, and a method for extracting tiles from all levels is used to extract the urothelium tiles to be used for grading. **Diagnostic prediction)** The urothelium tiles are fed to a diagnostic model, which outputs a probabilistic score for the two classes, low- and high-grade carcinoma. The system will output a WHO04 grade heatmap and a slide-level WHO04 prediction.

The proposed method was demonstrated by displaying heatmaps of each tissue class on unseen WSIs. In Paper 2, we leveraged on pre-trained models and domain adaptation to better adapt our models to the small labeled dataset. We extracted the dataset at three magnification levels (25x, 100x, 400x) to be able to utilize multiscale models. Three novel architectures were proposed, referred to as MONO-, DI- and TRI-scale models. Furthermore, new and novel tissue segmentation maps were implemented to demonstrate the methods on WSIs. Lastly, in Paper 3, we utilized semi-supervised methods on the best-performing model from Paper 2. A clustering approach and a probability approach were experimented on to improve the classification of the different tissue classes.

In Paper 4, an automatic and parametric method for extracting tiles in multilevel gigapixel was proposed. The method is parameterized and, as such, repeatable, reproducible, and easy to report by reporting a few parameters. First, the full WSI dataset was segmented into all tissue classes using the best model from Paper 2. Afterward, the methods described in Paper 4 were used to extract urothelium tiles from all three magnification levels to create a diagnosis dataset. Finally, in Paper 5, a model for predicting low- and high-grade carcinoma was proposed. The method correctly graded 45 of the 50 WSIs in the test set, achieving an accuracy of 90%, and the method was further demonstrated by creating heatmaps

on 14 WSIs annotated into low- and high-grade carcinoma regions by a pathologist, achieving a weighted average F1-score of 83%.

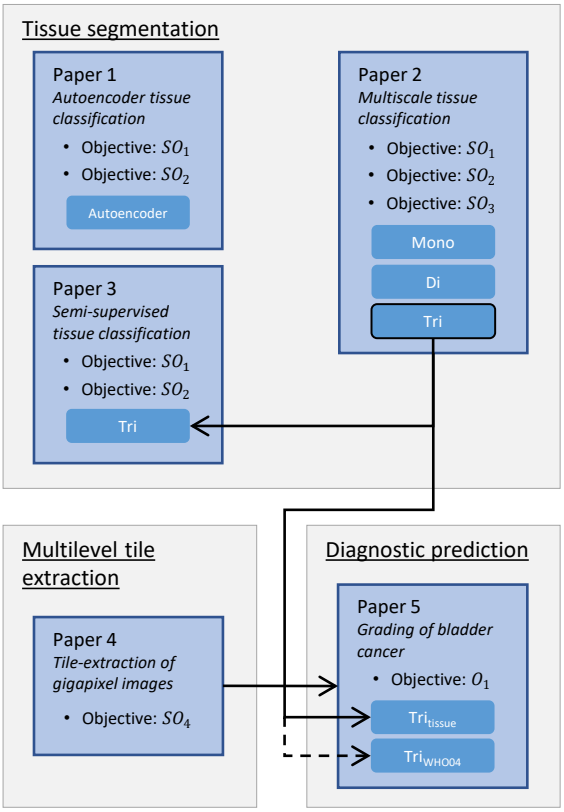


Figure 1.3: An overview of the contributions and the relationship between them. The thesis objectives and the main model used are shown for each paper. The first three papers are devoted to tissue segmentation and explore the objectives SO_1, SO_2, and SO_3. Paper 4 is for multilevel tile extraction and investigates the objective SO_4. Paper 5 is for diagnosis prediction, the objective O_1. It leverages the TRI-model from paper 2 and methods from Paper 4.

1.6 Thesis outline

The remaining content in this thesis is organized as follows: Chapters 2 and 3 will provide an overview of the relevant background theory used in this thesis for medical and technical background, respectively. In Chapter 4, the data material used in the thesis is presented, and it is describes how

the different datasets are defined. Chapter 5 will give a synopsis of Paper 1, 2, and 3, relevant to the topic of tissue segmentation, while Chapter 6 will summarize the contribution of Paper 4 and the work on multilevel tile extraction, and Chapter 7 will present Paper 5 related to the topic of diagnostic prediction. Chapter 8 contains the discussion and conclusion of the thesis. Finally, the five papers are reformatted and presented in Chapters 9 to 13.

Chapter 2

Medical background

In this chapter, an introduction to bladder cancer is given. Then, an overview of bladder cancer diagnosis is given, and, in the end, the epidemiology, incidence, and mortality of bladder cancer are given.

2.1 Bladder cancer

Cancer of the bladder is known as bladder cancer and is a disease in which abnormal cells multiply without control and form tumors in the urinary bladder. Tumors may be found anywhere within the bladder but are most common along the lateral walls [81]. Bladder cancer requires an intensive treatment and follow-up plan, which results in it being one of the cancer types with the highest lifetime treatment cost per patient [12, 111].

The urinary bladder is a hollow muscular organ that functions as a reservoir for storage of urine. The urine comes from the kidney, enters the bladder via the ureters, and exits the bladder via the urethra. The inside of the bladder is lined with muscle tissue that stretches to hold the urine. A cross-section of the urinary bladder is depicted in Figure 2.1, showing the different tissue types making up the bladder wall. The bladder lining consists of the urothelium and acts as a membrane. Below the urothelium is the connective tissue made up of stroma tissue, followed by a layer of muscle tissue and a layer consisting of fat.

In the same figure, example tumors of different T-stages are also shown. The Tumor Node Metastasis (TNM) classification system defines the cancer stage depending on how far the tumor has spread into the surrounding tissue. The tumor stage classification system ranges from CIS to T4, where T4 is the most invasive and has the worst prognosis. In its earliest stages (CIS, Ta, T1), the tumor is confined to the bladder lining or stroma tissue

and has not invaded the muscle tissue. These stages are known as non-muscle-invasive bladder cancer (NMIBC) and are easier to treat. Whereas for stages T2 to T4, the tumor has invaded the muscle wall and is referred to as muscle-invasive bladder cancer (MIBC). This is a severe condition, and a cystectomy is often required, i.e., removal of the bladder. All patients in this thesis have either cancer stage Ta or T1, meaning they have NMIBC.

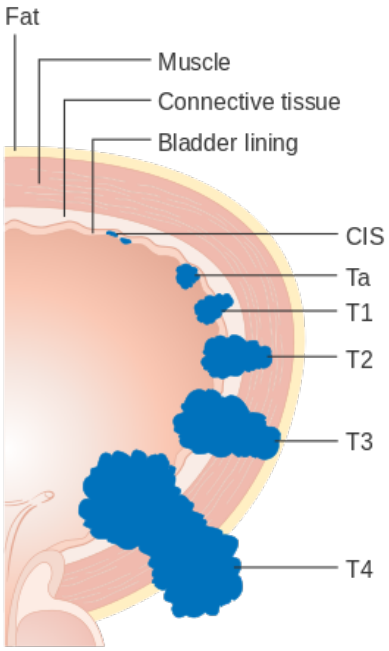


Figure 2.1: A cross-section of the urinary bladder showing the T-stages of bladder cancer, and how the cancer tumor infiltrates the nearby tissue. Image by Cancer Research UK, licensed under the Creative Commons BY-SA 4.0 license [125].

For patients diagnosed with NMIBC, the tumor is usually removed through transurethral resection of bladder tumor (TURBT). The removed tissue contains both atypical urothelial from the tumor and stroma, but can also contain smooth muscle from the bladder wall, normal urothelium from surrounding mucosa and blood. During the procedure, parts of the tissue can get both physical- and heating damage, for example, in terms of heating damage induced by laser or electrically heated wire loop, also called cauterization damage, or tearing of the tissue samples. Areas in the WSI with damaged tissue or blood will not be suitable for extracting

diagnostic and prognostic information, and a pathologist will ignore such regions during an examination.

For the purpose of grading NMIBC, urothelium is the most diagnostic relevant tissue. For staging, both urothelium and stroma, and particularly the border between them, is essential. The presence of muscle tissue also has importance for correct staging. However, cauterized tissue from the TURBT process, as well as areas containing blood, has no diagnostic relevance. Examples from each class are shown in Figure 2.2.

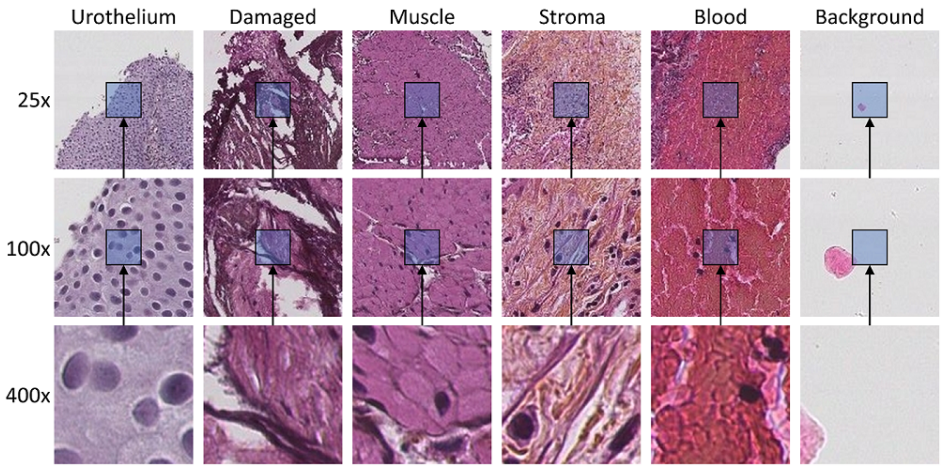


Figure 2.2: Example tiles of each class extracted at three magnification scales. Tiles at each scale are extracted from the same center pixel. The magnification scale is increased by a factor of four in each step, resulting in the tile covering 16 times as much area, even though they have the same size of 128x128 pixels.

The removed tissue is prepared, stained, and placed on a glass slide for analysis through a microscope. Examination of the tissue specimen is a process referred to as histopathology and is performed by a pathologist. It is a manual analysis that is very challenging and time-consuming. To aid the pathologists, different chemical dyes are used during the staining process. The stain creates contrast and emphasizes different aspects of the tissue, like immune cells or different tissue types. A variety of staining methods exists, depending on what features to highlight. All WSIs used in this work are stained either using haematoxylin, eosin, and saffron (HES), or haematoxylin and eosin (H&E). The haematoxylin will stain the cell nuclei in a purple-blue color, the eosin will stain the extracellular matrix and cytoplasm in a pink color, giving the WSIs its recognizable look, and saffron is used to distinguish fibers of collagen [24, 26].

The most prominent risk factors for bladder cancer are cigarette smoking and occupational exposure to chemicals [106]. With bladder cancer, as with other cancer types, there is a risk of metastasis, where the cancer is spreading to nearby lymph nodes or other organs. However, the primary focus of this thesis is the classification and grading of urothelial carcinoma.

2.2 Diagnosis

In histopathological diagnostics, pathologists use grading and staging to describe the tumor. These parameters are used to stratify patients into risk groups and tailor a suitable treatment and follow-up plan.

A histological image will reveal specific diagnostic information at different resolutions, and a pathologist will integrate information across several magnification levels before reaching a decision. High magnification (400x) will reveal cytological features like cell size and shape, mitosis, as well as cell nucleus characteristics as contour, size, and colorization (intensity and distribution). As you go down in magnification, you will get a broader field of view and show more context information from the surrounding tissue. At 100x, you can evaluate nucleolar polarity, and lower magnification (25x) will show global context information such as papillary architecture, outline, and border of the tissue, as well as color and texture.

After examination, the pathologist will document his or her findings in a pathology report. This report will include histological description and information about the grade and stage, and an estimate of the risk for recurrence and disease progression.

The tumor stage is important and is determined based on the size of the primary tumor, if it has invaded nearby tissue, and if so, how far it has spread into the surrounding tissue, as well as the number of primary tumors present. Pathologists use the TNM classification system to stage bladder cancer tumors, and example tumors for each stage are shown in Figure 2.1. The tumors may form papillary protrusions into the bladder lumen, solid nodules, or grow diffusely within the bladder wall. Approximately 70% of patients have NMIBC at first diagnosis [81], where the tumor has not invaded the muscle wall.

The grade of a tumor describes the differentiation state of the tumor cells under a microscope. Different cancer types have different grading scales, but in general, if the cancer cells are similar to that of healthy non-cancerous cells, the grade will be low, and the cancer will have a lower likelihood of

spreading. On the other hand, if the cells have a more abnormal appearance and are disorganized, the grade will be higher. The grade is generally based on the tissue architecture, nuclear arrangement, proliferation, and nuclear atypia. Each of these categories has several subcategories to describe the tumor in detail [81]. In [65], a set of 13 microscopic features are listed, which are examined to determine the final grade of the tumor.

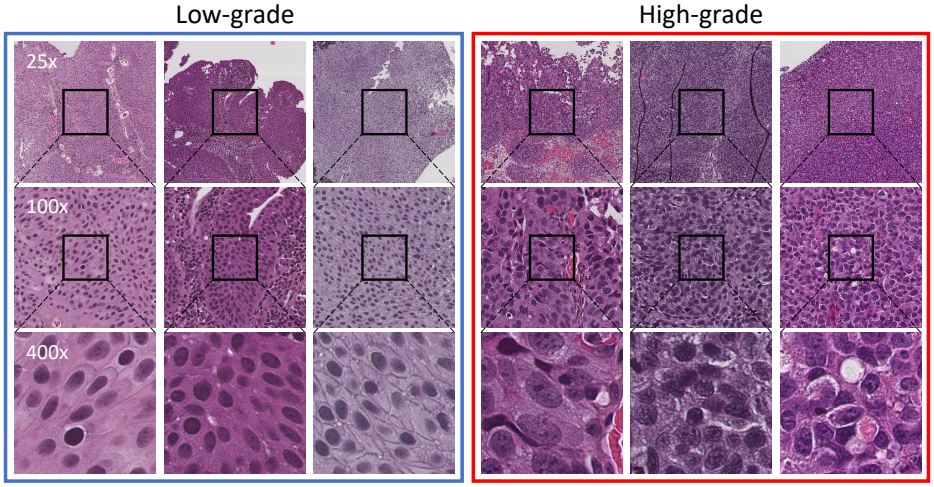


Figure 2.3: Examples of low-grade and high-grade tiles extracted from a WSI. The tiles are extracted from three magnification levels (25x, 100x, and 400x) and all have the same size of 256 × 256 pixels.

The World Health Organization (WHO) has proposed three grading systems for bladder cancer. The first grading system was introduced in 1973, referred to as WHO73, which is still somewhat used today. It consists of three categories, grade 1, grade 2, and grade 3, where grade 3 is the most severe state. A revised edition of the grading system was introduced in 2004 called WHO04, and further updated in 2016 as WHO16. In these versions, cases are split into low- and high-grade carcinoma. Some examples of low- and high-grade areas are shown in Figure 2.3. Grade 1 patients are referred to as low-grade patients, and grade 3 patients are high-grade patients. Patients diagnosed as grade 2, however, are now split into either the low- or high-grade case. This might seem like a minor change, but for a patient to be diagnosed as low- or high-grade may result in very different follow-up regimes and local treatment with potential adverse events. A patient falsely diagnosed as a high-risk patient is an example of unnecessary patient suffering by overtreatment, additional load on the health care system, and

extra cost. There is some correlation between the WHO73 and WHO04 systems, but they are not directly interchangeable, so both systems coexist [81]. The data material used in this paper was collected and diagnosed prior to 2016 and will therefore focus on the WHO04 grading system.

A WSI may contain regions of different grades, as well as regions of normal urothelium. It is usually assigned the worst grade present in the WSI as the final diagnosis.

2.3 Epidemiology

Bladder cancer is the 10^{th} most commonly diagnosed cancer worldwide, with an estimated 573 278 new cases and 212 536 deaths in 2020 [119]. Figure 2.4 and 2.5 shows an estimate of age-standardized incidence and mortality for bladder cancer in 2020.

It is well known that men are overrepresented when it comes to bladder cancer. Figure 2.6 shows the estimated incidence and mortality rates. The left-hand side of the plot shows the estimated numbers of incidents and mortalities for males and the right-hand side for females.

And finally, Figure 2.7 presents a diagram showing the most common cancer types for men in 2020. With 4.4% of the new cases falling into the bladder cancer category, it is the 6^{th} most common cancer among men.

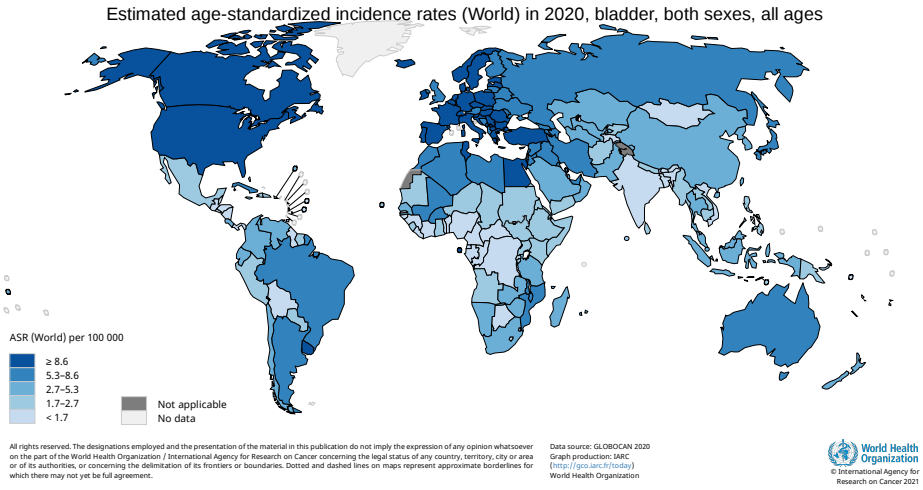


Figure 2.4: Estimated age-standardized incidence rate per 100 000 for both sexes in 2020. Reprinted from Global Cancer Observatory: Cancer Today. Public domain [38].

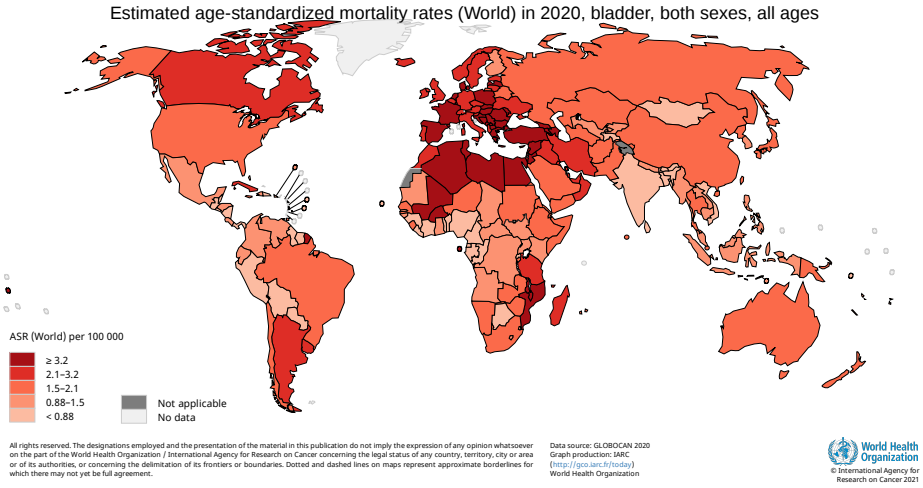


Figure 2.5: Estimated age-standardized mortality rate per 100 000 for both sexes in 2020. Reprinted from Global Cancer Observatory: Cancer Today. Public domain [38].

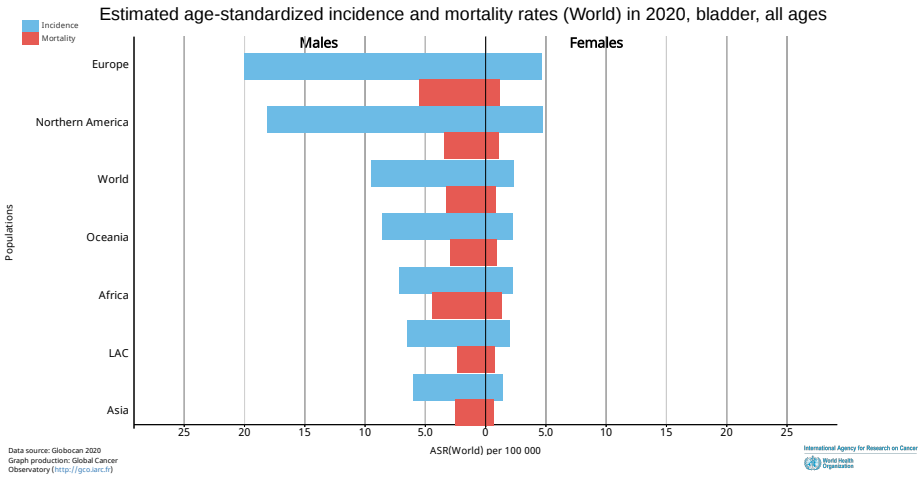


Figure 2.6: Estimated age-standardized incidence and mortality rates for bladder cancer in 2020. Rates for males are shown to the left and for females to the right. Reprinted from Global Cancer Observatory: Cancer Today. Public domain [38].

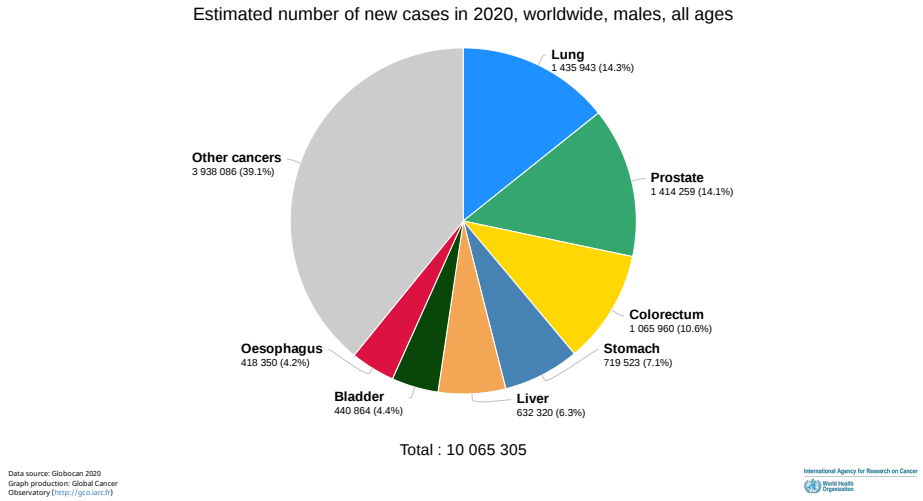


Figure 2.7: Estimated number of new cases of all cancer types in 2020, for males only. Bladder cancer accounts for 4.4% of the new cases, resulting in it being the 6^{th} most common cancer among men. Reprinted from Global Cancer Observatory: Cancer Today. Public domain [38].

Chapter 3

Technical background

In this chapter, an introduction to artificial intelligence is given, followed by an overview of the different building blocks in a neural network. Then, different evaluation metrics and how to distribute the datasets are presented. Finally, the most common learning techniques are discussed. The chapter focus on the techniques used in the thesis as the overall topic is far too wide to cover comprehensively in the scope of this thesis.

3.1 Artificial intelligence

Artificial intelligence is a multidisciplinary field of study involving neurobiology, information theory, cybernetics, statistics, computer science, and more. Artificial intelligence aims to train an intelligent agent to solve specific tasks based on the environment presented to the agent. The intelligent agent wants to optimize its success by taking the decisions with the highest probability. Machine learning is a subfield of artificial intelligence used to train an artificial intelligence system. First, an algorithm is trained on a dataset, and then the performance is measured against an independent test set. If the performance on the test set increases over time as the algorithm is trained on the training set, the algorithm is said to be learning. In essence, a dataset is a collection of samples, where each sample is a collection of features. In machine learning, we want to develop an algorithm to learn these features using the dataset.

A machine learning algorithm can be trained for many different tasks, for example, regression, classification, or segmentation. In regression, the predicted output will be a continuous value (e.g., predicting temperature, house pricing, stock market pricing.), whereas in classification, the prediction will be a discrete, categorical value. Examples of classification problems can be an email spam filter (spam vs. not spam), classifying

images of animals (cat vs. dog), or predicting diagnosis for a disease (cancer vs. non-cancer). Segmentation is a computer vision problem, where the core problem is understanding the scene. From the input image, the segmentation algorithm will classify each pixel into one of the predetermined classes. Machine learning can be used in a wide range of other tasks as well, such as machine translation, anomaly detection, synthesis, clustering, imputation of missing data, or denoising, but those aspects will not be discussed here.

To create a successful machine learning system, three elements are required; a sufficiently large set of training data, computational power to process the data, and an algorithm that learns from the data. Nowadays, advances are still made on each of the three required elements. Manufacturers of graphical processing units (GPU) are pushing their limits and developing new and more powerful units each year. Simultaneous, cloud computing has grown in both accessibility and popularity, allowing users almost unlimited computational power at the expense of cost. Large, open-source datasets are growing in both numbers and size. There is also a trend to arrange competitions, where the chairholders have gathered a large, often labeled, dataset, and participants are encouraged to develop algorithms and compete against each other. This also takes place here in Norway, where recently the Norwegian Artificial Intelligence Research Consortium (NORA) announced MedAI, a medical image segmentation competition to segment polyps in images taken from endoscopies [91]. And finally, researchers are developing new methods and algorithms in record speed within artificial intelligence. Benjamens et al. state that "*the number of life science papers describing AI/ML rose from 596 in 2010 to 12 422 in 2019*" [9].

During the past decade, AI-related methods have also bridged the gap from research over to finalized products. For example, a database of FDA-approved AI/ML-based devices is presented in [9], and an up-to-date database is maintained at the website [124]. The year 2010 marked the first year where the FDA approved an AI/ML-based device, and the number of such approvals has since only increased, as seen in Figure 3.1.

3.2 Deep learning networks

This chapter introduces neural networks and some common techniques and models used in deep learning.

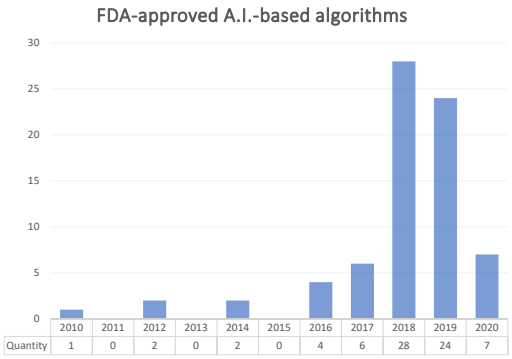


Figure 3.1: The number of FDA-approved AI/ML-devices. Data collected from [124].

3.2.1 Neural networks

Artificial neural networks draw inspiration from the human brain, where the biological neuron is modeled into an artificial neuron. A depiction of an artificial neuron is shown in Figure 3.2. It consists of multiple inputs x_i and weights w_i, where $i = 1, \ldots, m$. The input x_0 is usually set equal to 1, and then the weight w_0 is used for the bias term, often written as b. A weighted sum between inputs and weights are computed, and the bias term is added, $b + \sum_{i=1}^{m} x_i w_i$. The resulting sum is fed through an activation function that generates a single output. Each input to the neuron has a weight associated with it, which can strengthen or weaken the signal. The bias value is used to shift the weighted sum. The activation function defines the neuron's output and determines if the neuron should be activated or not. A wide range of activation functions exists, all with different properties [66].

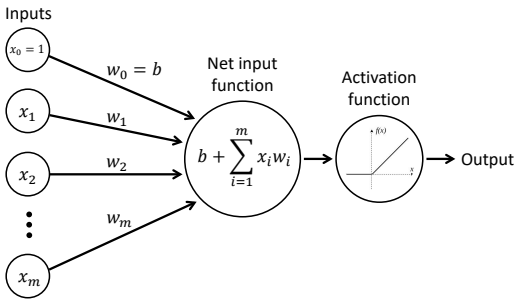


Figure 3.2: Example of an artificial neuron. It computes a weighted sum from the inputs and computes a single output.

Similar to how the brain links neurons together in a network, an artificial neural network also consists of multiple neurons linked together and organized layer by layer. An example of a simple neural network is depicted in Figure 3.3. The network consists of an input and output layer, and all layers between the input and output are referred to as the hidden layers. Because each neuron is connected to all neurons in the following layer, these layers are also sometimes referred to as fully-connected layers. And a network consisting of only fully-connected layers are sometimes referred to as a fully-connected neural network (FCNN).

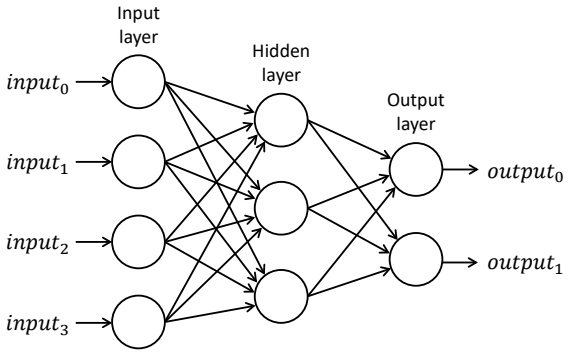


Figure 3.3: Example of an artificial neural network. Each node consists of an artificial neuron as depicted in Figure 3.2.

It is possible to add more hidden layers, and the number of layers in a network determines the depth of the model. Over the past decade, the number of layers in models has increased rapidly, and models have become deeper and deeper. This has led to the introduction of the term deep learning, referring to a machine learning task solved using a deep model.

The weights and bias values are usually initialized randomly from a truncated Gaussian distribution before training starts. Training a neural network is an iterative optimization problem, usually based on stochastic gradient descent (SGD), and requires a loss function to quantify the error in the model's predictions. The optimization algorithm uses an algorithm called backpropagation for calculating the gradient of the loss function with respect to the weights and bias values. The optimization algorithm then updates the weights and biases by minimizing the loss from the loss function. Multiple loss functions exist, but for the training of neural networks, the most common functions are cross-entropy and mean squared error (MSE) [66].

3.2.2 Autoencoders

An autoencoder is a neural network with a specific network design, as shown in Figure 3.4. The goal of an autoencoder is to learn how to reconstruct the input on the output. The network typically has a bottleneck structure, where the middle part is smaller than the input and output. It is one consistent network, but it is common to refer to the first half of the network as the encoder, and the latter part as the decoder, as shown in Figure 3.4. The overall structure of the encoder decreases in size, forcing the model to discard redundant features and learn the features that are important for reconstruction. This makes autoencoders great for learning feature extractors without the need for detailed labels. The smallest layer in the network, often called the bottleneck layer or latent vector, is a latent feature representation containing a code describing the input. The term latent vector means that the stored values are hidden and not directly observable, hence the need for a decoder to reveal the stored information. The role of the decoder network is to decode the code stored in the latent vector and thus reconstruct the input as closely as possible. The size of the layers in the decoder is usually the same as in the encoder but in reversed order.

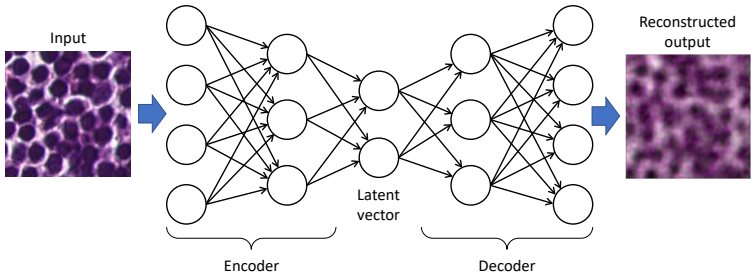


Figure 3.4: Typical structure of an autoencoder model. The input is transformed into a latent vector using the encoder part of the network and then reconstructed by the decoder network.

To train an autoencoder, the MSE loss function is utilized. This function measures the reduced mean of the squared difference between the output and input with the following expression:

$$Loss_{MSE} = \frac{1}{NM} \sum_{i}^{N} \sum_{j}^{M} (Output_{ij} - Input_{ij})^2 \tag{3.1}$$

One of the main advantages of the autoencoder model is that the loss

function does not require an input label. Because of this, the autoencoder is often used in unsupervised learning.

3.2.3 Convolutional neural networks

A convolutional neural network (CNN) is a neural network where at least one of the layers of the model consists of a convolutional layer. These convolutional layers use the mathematical operation convolution between its input and a kernel consisting of weights. It is these weights in the kernel that is adjusted during training. Convolutional networks work excellent on data such as images, but are also employed on other data types like time-series data.

Some of the advantages of using a convolutional network are parameter sharing and sparsity of connections. For parameter sharing, the convolutional layer relies on the same filter kernel, which strides across the entire input. Since the filter kernel is much smaller than the input, each output value depends only on a small number of inputs, resulting in the sparsity of the connections. A result of both parameter sharing and the sparsity of connection is fewer parameters in the network, making the convolutional network more memory efficient than regular neural networks. In addition, because the filter is shifted across the entire image, specific features can be detected at any location in the image, resulting in one of the properties of convolutional networks known as shift-invariance.

Pooling layers are often used in CNN models, usually added after a convolutional layer or a group of convolutional layers. Unlike the convolutional layers, the pooling layers do not contain any parameters; instead, it is an operation that downsamples its input. The pooling layer slides a small region across the input, usually a 2×2 pixel region with a stride of 2 pixels, and applies the pooling operation at each location. The most common operation types are average pooling and maximum pooling. The average pooling operation computes the average value of the samples within the region, and max pooling selects the maximum value from the region. The pooling layers are used in convolutional networks because they reduce the dimensions of the feature maps, which again helps with reducing the total number of parameters in the model. Another benefit is that the pooling layers make the model invariant to small translations. Because the pooling layers aggregate the features within a small region, translations within this region would often result in the same aggregated values by the pooling operation.

3.2.4 VGG16 convolutional neural network

VGG16 is the name of a CNN architecture proposed by K. Simonyan and A. Zisserman [113]. The architecture is depicted in Figure 3.5 and consists of five convolutional blocks followed by three fully-connected layers. The convolutional blocks consist of two or three convolutional layers and a max-pooling layer.

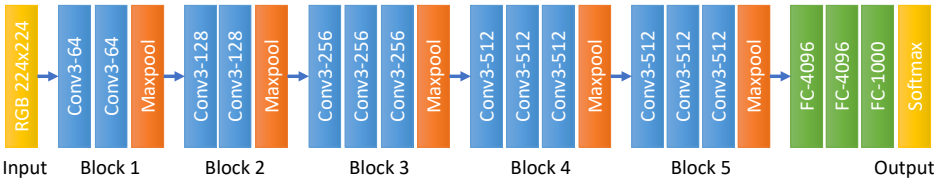


Figure 3.5: The VGG16 architecture. The convolutional layers indicate the receptive field size and number of channels. The fully-connected (FC) layers indicate the number of neurons in the layer. Based on architecture description in [113].

The name VGG16 stems from the group's name, Visual Geometry Group from the University of Oxford, and the number 16 refer to the network containing 16 trainable layers. The max-pooling layers do not contain any parameters and are therefore not considered a trainable layer. The VGG team submitted their proposed model to the 2014 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) and won first and second place in localization and classification tasks. After the competition, the team further improved the model and shared their model with pre-trained weights. The model is included in most deep learning frameworks, such as Keras, Tensorflow, and PyTorch, making it easy for researchers to download and use the model. Because of the model's availability, it has become a very popular model to use, and by November 2021, the paper currently has over 65 000 citations on Google Scholar.

3.2.5 Regularization techniques

When training a deep learning model, there is always a risk of overtraining the model. That is, to optimize the weights on the training dataset to such an extent that the model's generalization is reduced. As a result, the accuracy score would be relatively high if the model would be evaluated on the training dataset. However, if assessed on an independent test set, the score would be much worse. This undesired effect is referred to as overfitting on the training data and should be avoided. Large models trained on small

datasets are more prone to overfitting. To lower the overfitting problem, regularization techniques are usually implemented.

Early stopping

The problem described above can arises when training for too long. However, the problem would be resolved if model weights from earlier training were restored into the model.

Early stopping is a technique where the model is evaluated on a validation dataset after each epoch of training. The model's weights are not updated during this step; only the validation accuracy or loss is stored. Training of the model then continues until the performance on the validation set does not improve for a predetermined number of epochs. Once training is terminated, the weights from the best-performing epoch are restored.

Dropout layers

Dropout layers are placed between other layers in a neural network. They have no weights associated with them. Instead, nodes in the dropout layer are randomly set to zero during training. For each training step, a new set of nodes are randomly dropped out. This has a regularization effect on the network, as the connectivity between the layers is altered, and smaller subnets in the network emerge. The dropout effect is disabled during evaluation on the validation or test sets.

3.3 Evaluation metrics

When training a deep learning model, there is a large unknown search space of hyperparameters and other choices a data architect needs to take. These choices range from which preprocessing technique to implement, how many layers are in the model, what type of layers, as well as post-processing steps. For hyperparameters, we need to choose an appropriate learning rate, batch size, dropout rate, number of neurons in each layer, which optimizer to train the model, to name a few. How we arrange our datasets is also important. How we distribute the data into training, validation, and test sets, or if we choose to use cross-validation or other ways to split our data.

These are just some of the more important choices to make when designing a deep learning system for a specific task and set of data. Unfortunately,

one of the challenges with machine learning, in general, is that there is no obvious way to determine the optimal parameter for each of these choices. Instead, a set of models are often trained on different parameters, and then the performance of each model is compared with each other using a validation set. Model architecture tuning and hyperparameter optimization is an iterative process. A set of default parameters are trained and evaluated, and then the evaluation metric will guide which actions to take.

A set of different metrics are used to measure a model's performance and will be explained here.

3.3.1 Confusion matrix

A confusion matrix summarizes the model's prediction into different classes, and is suitable for classification tasks. An example confusion matrix with three classes is shown in Figure 3.6. The rows represent the true classes, and the columns represent the predicted classes. In some situations, the number of predicted samples is shown for each class, as in the example confusion matrix in Figure 3.6. This can be useful as it shows exactly how many samples are predicted correctly and incorrectly for each class. However, if the classes are imbalanced, it may, in some cases, be more beneficial to normalize the predictions in the confusion matrix.

Because the confusion matrix contains a large set of numbers, it may be difficult and time-consuming to compare the performance of different models. Hence, it is often desirable to aggregate, or extract, a few metrics indicating the performance of the models. Many different metrics can be derived from the confusion matrix, from which some will be presented here.

For a classification problem with n classes, a $n \times n$ confusion matrix is constructed. Each element in the confusion matrix can be referenced as $\text{cell}_{\text{i,j}}$, as seen in Figure 3.6.

3.3.2 TP, FP, FN, and TN

In confusion matrices, the terms *positive* and *negative* are often used and are usually linked to an outcome. For instance, positive may refer to a patient with a disease and negative to a patient without the disease. Based on the values of true and predicted positive and negative cases, four metrics can be extracted. First, the true positive (TP) value is the number of samples where the true class and the predicted class are the same; i.e., the

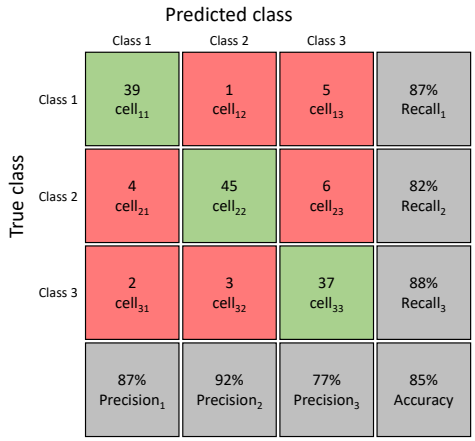


Figure 3.6: An example confusion matrix with $n = 3$ classes. The green boxes indicate correct predictions, and the red boxes are incorrect predictions. The gray boxes are different aggregated evaluation metrics used to describe the performance of the current model.

TP value refers to the number of correct predictions. The false positive (FP) value is the number of samples that belong to the negative class but is wrongly predicted as the positive class, hence the name false positive. The false negative (FN) value is the number of samples that belong to the positive class but are wrongly predicted as one of the negative classes, hence the name false negative. Finally, the true negative (TN) value is the number of negative samples correctly predicted as negative, hence the name true negative.

In the binary case ($n = 2$), values for TP, FP, FN, and TN are computed as described above. However, in a multiclass setting ($n \geq 3$), the values for TP, FP, FN, and TN must be computed for each class [122]. The following equations will compute the values for TP, FP, FN, and TN for class c:

$$\mathrm{TP}_c = cell_{c,c} \tag{3.2}$$

$$\mathrm{FP}_c = (\sum_{P=1}^{n} cell_{P,c}) - \mathrm{TP}_c \tag{3.3}$$

$$\mathrm{FN}_c = (\sum_{Q=1}^{n} cell_{c,Q}) - \mathrm{TP}_c \tag{3.4}$$

$$\mathrm{TN}_c = (\sum_{P=1}^{n}\sum_{Q=1}^{n} cell_{P,Q}) - \mathrm{TP}_c - \mathrm{FN}_c - \mathrm{FP}_c \tag{3.5}$$

3.3.3 Total population

The total population is the sum of all elements in the confusion matrix, corresponding to all instances.

$$total\ population = \sum_{P=1}^{n}\sum_{Q=1}^{n} cell_{P,Q} \tag{3.6}$$

3.3.4 Accuracy

Accuracy is an overall measure of the model's performance. It is the proportion of correctly predicted samples and computed as follows:

$$Accuracy = \frac{\sum_{P=1}^{n} \mathrm{TP}_P}{total\ population} \tag{3.7}$$

Accuracy is a single metric of the model's overall performance, making it easy to compare several models and is often used in result tables. However, if the classes are imbalanced, it may not give an accurate evaluation of the model.

3.3.5 Precision

Precision is the proportion of positive predictions that are positive. For example, in Figure 3.6, the precision for class 2 is 92%, which means that of all the samples the model predicted as class 2, 92% of them belongs to the true class 2. Precision is sometimes referred to as positive predictive value (PPV). Precision needs to be computed per class. Expression for precision for class c is:

$$Precision_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c} \tag{3.8}$$

3.3.6 Recall

Recall is the proportion of positive samples that are correctly predicted as positive. For example, from Figure 3.6, the $recall_1$ of 87% corresponds to that 87% of the true class 1 instances are predicted as class 1. Recall may also sometimes be referred to as sensitivity or True Positive Rate (TPR). Recall needs to be computed per class. Expression for recall for class c is:

$$Recall_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c} \tag{3.9}$$

3.3.7 F1-score

The F1-score is the harmonic mean between precision and recall, computed as follows:

$$\mathrm{F1}_c = 2 \cdot \frac{Precision_c \cdot Recall_c}{Precision_c + Recall_c} = \frac{2 \cdot \mathrm{TP}_c}{2 \cdot \mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c} \tag{3.10}$$

F1-score is a popular metric used in many applications, such as information retrieval, machine learning, and natural language processing. It is also well suited to use with imbalanced data.

3.4 Data distribution and augmentation

Machine learning algorithms are, in essence, algorithms that will learn from experience. This experience is usually gathered from a dataset that is presented to the algorithm. Thus, the dataset itself is a vital part of any machine learning system. How we use the data at our disposal and train our model may impact the resulting model. This section look at some common ways to process the datasets for deep learning models.

3.4.1 Distribution and splitting

Train, validation, and test

The total dataset is often split into three parts: training, validation, and test set. The training set is used to optimize the parameters of the model during training. The validation set is never used for weight adjustments but is used

for selecting hyperparameters and evaluating different architectures. The final model is then evaluated on the test set for the final result. Depending on the size of the training dataset, it can sometimes be helpful to re-train the final model architecture with the chosen hyperparameter settings on the combined training and validation dataset before evaluation on the test set.

K-fold cross-validation

In k-fold cross-validation, the dataset is first split into two parts: the data used for cross-validation and a test set. The samples from the cross-validation part are then further divided into k number of folds of approximately equal size. Example of a 5-fold cross-validation setup is depicted in Figure 3.7. The model is then trained on $k-1$ folds and validated on 1 of the folds. This process is repeated for k iterations, with a different fold used for validation for each iteration [103, 144]. Once all iterations are completed, we are left with k models. Therefore, it is necessary to re-train the model architecture with the hyperparameter settings on the combined data from all k folds.

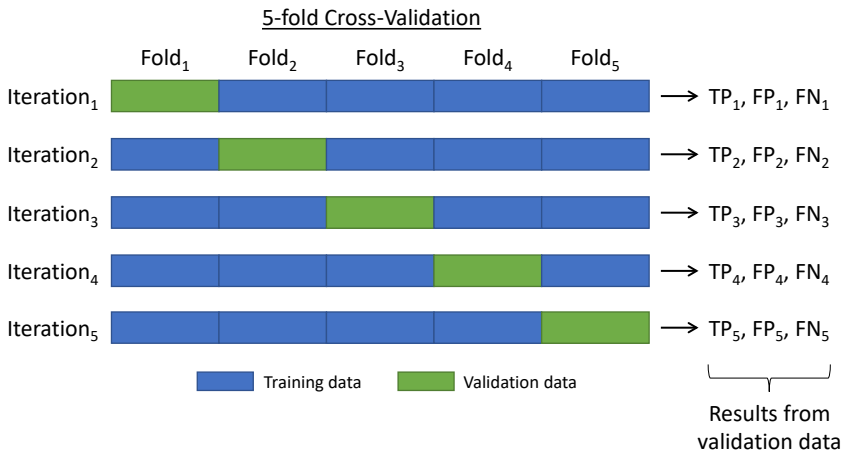


Figure 3.7: Example of a 5-fold cross-validation setup. The available data is split into five equal folds. For each iteration, a model is trained on the blue folds and validated on the green fold. Results are saved for each iteration and used to judge the performance of the model.

One of the main advantages with this setup, is that each sample in the cross-validation data is used as both training and validation. This

makes it useful in scenarios where only a small amount of data is available. A downside, however, is the requirement of training k models. This is especially challenging in a deep learning setting, where each model can take several days to train.

An index value i is used to keep track of the different iterations, where $i = 1, \ldots, k$. After a model has completed training on iteration$_i$, the model is evaluated on the validation data in fold$_i$, and values for TP_i, FP_i, and FN_i are computed. To compute an overall F1-score for all the iterations combined, there are two approaches referred to as macro- and micro-averaging.

For macro-averaging, the values for TP_i, FP_i, and FN_i are used to compute the $\mathrm{F1}_i$-score for each respective iteration according to Equation 3.11. Then, using the $\mathrm{F1}_i$-score from all iterations, the macro-average F1-score is calculated with Equation 3.12.

$$\mathrm{F1}_i = \frac{2 \cdot \mathrm{TP}_i}{2 \cdot \mathrm{TP}_i + \mathrm{FP}_i + \mathrm{FN}_i} \tag{3.11}$$

$$\mathrm{F1}_{macro-avg} = \frac{1}{k} \sum_{i=1}^{k} \mathrm{F1}_i \tag{3.12}$$

For micro-averaging, the values for TP_i, FP_i, and FN_i are summarized for each fold using Equations 3.13, 3.14, and 3.15. Then, the micro-average F1-score is computed using Equation 3.16. According to Forman and Scholz [41], this is the recommended way of computing the F1-score in a cross-validation setup and will produce an unbiased evaluation score.

$$\mathrm{TP}_{tot} = \sum_{i=1}^{k} \mathrm{TP}_i \tag{3.13}$$

$$\mathrm{FP}_{tot} = \sum_{i=1}^{k} \mathrm{FP}_i \tag{3.14}$$

$$\mathrm{FN}_{tot} = \sum_{i=1}^{k} \mathrm{FN}_i \tag{3.15}$$

$$\mathrm{F1}_{micro-avg} = \frac{2 \cdot \mathrm{TP}_{tot}}{2 \cdot \mathrm{TP}_{tot} + \mathrm{FP}_{tot} + \mathrm{FN}_{tot}} \tag{3.16}$$

Stratified k-fold cross-validation adds one additional requirement to the setup. Each fold must contain about the same number of samples from each class. This ensures that each fold is a good representation of the whole dataset [41].

3.4.2 Data augmentation

The aim of data augmentation techniques is to utilize existing data and generate synthetic data in a low-cost manner, usually automatically and without human intervention. Since the augmented data is based on data from the actual dataset, it is representative of the specific dataset. The augmented data will typically be used in training and not during validation or testing. Having access to a larger training dataset can potentially help produce more robust models. It is also possible to only augment data from a few of the classes in the dataset to help balance an imbalanced set.

For image data, the most common augmentation technique consists of rotating and mirroring. However, it can only be applied to rotationally invariant data, that is, images where the semantic meaning or value does not change when arbitrarily rotated. If an image does not change its class after rotation, the original label for the image can be reused for the augmented image. It is possible to rotate the image with an arbitrary number of degrees, but then the resulting image is not square anymore, making it necessary to crop the image before feeding it to a neural network. By limiting the rotations to factors of 90 degrees and flipping horizontal and vertically, it is possible to augment an input image in eight unique ways.

A range of other augmentation techniques also exist. For example, in color augmentation, the RGB or HSV values of the image are slightly altered, changing the image's color. Random cropping is a technique where random sub-images are extracted from a larger image, and all sub-images represent the same category and therefore use the same label. It is also possible to add random noise to the input during training in a technique known as noise perturbation. For all augmentation techniques, care must be taken to not alter the features of the image to such an extent that it no longer represents its attached label.

3.5 Learning techniques

There are many techniques and methods used to train a model on a specific dataset. Often, the type of labels available for the dataset, or the absence of labels, influence the choice of method used to train the model. Some of the most common types will be presented here.

3.5.1 Supervised learning

In a supervised learning setting, we want to train a model on a dataset consisting of pairs of input samples and labels. The input set X contains a large collection of samples x, and a corresponding set Y contains one label y per input sample. There must exist some relationship between a sample x and its target label y, represented with the mapping function $y = f(x)$. The goal is to train a model to approximate this mapping function as $\hat{y} = f_{model}(x)$, and to use it in the future for classification of previously unseen input samples. Supervised learning is the most widely used learning algorithm to train neural networks and deep learning models.

During training, the labels in Y are compared with the model's prediction. Then, a loss function is used to calculate a distance measure between the model's prediction and the target. This distance, often referred to as error or loss, is then used to optimize the model's weights so that the distance decreases over time. The name *supervised* stems from the fact that the learning algorithm is given a set of ground truth labels by a supervisor, for example, a human-annotated label.

3.5.2 Unsupervised learning

In unsupervised learning, the goal is to find compact representations, clusters, and groups in the data without having any knowledge or labels on a training set. Therefore, unsupervised learning is harder to solve than supervised learning due to the lack of labels. In supervised learning, we can present the model with an image, and the target label will dictate what features within the image we want the model to learn. However, in unsupervised learning, we do not have such a luxury. Instead, we must alter the model's architecture and use constraints to force the model to solve the desired task.

Unsupervised learning can be used to solve different problems, such as dimensionality reduction and clustering. In dimensionality reduction, we

want to learn a model which can transform the input data into a new representation. Usually, there is some constraint in the model, forcing it to produce a simpler representation of the input. These new representations are often called low-dimensional representation, or sparse representation [44]. During learning, we want to preserve as much information regarding the input sample x as possible but still represent it in a simpler way. This is useful, as the new representation may be more accessible than x itself and give insight into the structure of the underlying data. In clustering, an algorithm will group similar input samples into categories based on a given similarity measure. The model will use the structure and pattern of the samples to group them.

There exist many popular probabilistic methods for unsupervised learning, such as k-means clustering, hierarchical cluster analysis (HCA), principal component analysis (PCA), and singular value decomposition (SVD). In addition, there are also unsupervised methods based on neural networks, e.g., restricted Boltzmann machine (RBM) and autoencoders. There are many applications of unsupervised learning, and some examples of usage ranges from content-based fake news identification [54], learning of probably symmetric deformable 3D objects from images [141], identify phases and phases transitions of many-body systems [135], anomaly detection [43], to unsupervised image segmentation [152] and image denoising [30].

3.5.3 Semi-supervised learning

Semi-supervised learning sits between supervised and unsupervised learning. It has access to a limited set of labels to optimize its parameters, but the majority of the training comes from unlabeled inputs. Semi-supervised approaches are usually pursued when the available labeled data is too small for a purely supervised learning setup. I.e., the model's performance will increase by incorporating the unlabeled data [93].

A study by Ligthart et al. [72] investigated the effectiveness of semi-supervised learning methods for opinion spam classification. They conclude that the "*self-training algorithm can outperform traditional supervised classification methods when limited labeled data is available*" and continue to state that "*the proposed semi-supervised approaches can mitigate labeling efforts while retaining high-performance*" [72].

There exist many different semi-supervised learning methods. One of these is called self-training (self-labeling or self-teaching is also used),

where a model is first trained in a supervised manner on a labeled dataset. Afterward, the trained model is used to annotate unlabeled data. This automatically generated labeled data can then be used as training data. This is an iterative process that may be repeated several times.

Another example of an iterative learning process is online human-assisted machine learning, or sometimes human-aided learning. First, a model is trained through unsupervised or semi-supervised learning techniques, followed by automatic annotation of unlabeled data. Next, a human will manually examine the model's prediction and correct the misclassified samples. If this were, for example, a segmentation task, the model would predict a ROI consisting of thousands of pixels. The human, however, may be able to classify the entire ROI in a short amount of time, therefore correcting a large number of data at once. Thus, human-assisted learning has the potential as an efficient method of generating a large dataset, especially in fields requiring expert input, like medical images.

Villamizar et al. built a classifier that progressively learned a face and object detection algorithm. During training, human intervention was used to assist the learning by discarding false-positive training samples [130].

3.5.4 Weakly-supervised learning

As mentioned, it is not always feasible getting enough labels of sufficient quality. An alternative to unsupervised learning is weakly supervised learning. This technique, sometimes referred to as weak supervision, uses labels of lower quality to train the models. These low-quality labels are cheaper and more efficient to produce but are imprecise or inaccurate in nature.

The labels may be sourced through crowdsourcing, where a large group of non-experts will collectively sample a large dataset. However, because there is little control over the labeling process and the use of non-experts, the labels may be inaccurate.

Another source of weak labels is diagnostic information and patient follow-up data. For example, a pathologist will grade and stage WSIs, but will usually not provide any localization data with the diagnosis. A pathologist will also record follow-up data such as recurrence and disease progression. However, these outcomes are not linked to a specific part of the WSIs. Using the diagnostic information and follow-up data on all parts of a WSI may not be completely accurate, and the labels for the extracted tiles would therefore be considered a weak label.

3.5.5 Domain adaptation

Domain adaptation, also known as transfer learning, is a technique that utilizes knowledge from one domain and transfers it to another domain. It is a very powerful method, easy to implement, and provides good results; and is thus a popular technique.

An overview of the domain adaptation setup is shown in Figure 3.8. First, a model is trained on a large dataset from one knowledge domain, for example, the VGG16 model from Chapter 3.2.4. The model is trained on the dataset to solve a specific task referred to as Task A. Next, we remove the last layers of the model and substitute them with new layers with random initialization. The model is then trained on a new dataset from another domain. The updated model is trained on the new dataset to solve a different task referred to as Task B. Training of the initial model on Task A is referred to as pre-training, while training the updated model on Task B is called fine-tuning. For transfer learning to work efficiently, the input datatype for tasks A and B must be the same. It is usually implemented in a situation where there exists a large labeled dataset for Task A but only a small limited dataset for Task B.

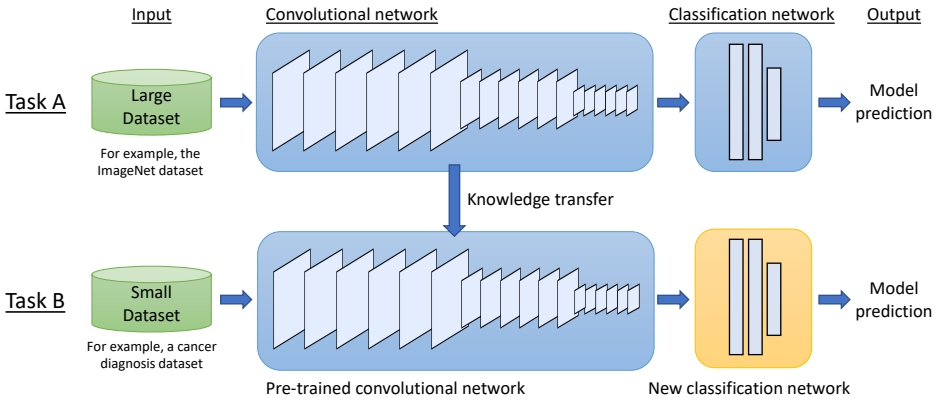


Figure 3.8: Overview of domain adaptation setup. A model is trained on a large dataset to solve Task A. Then, the knowledge is transferred to another domain, the last layers of the model are discarded and replaced with new layers. The new model is then fine-tuned on Task B.

Because the majority of the parameters in the model are pre-trained, training time is reduced during fine-tuning. In addition, when compared to training a model on Task B from scratch, it is common to see higher starting performance (i.e., the accuracy of the model on Task B is higher

at the start of training) and higher final accuracy on Task B when using transfer learning [92].

During fine-tuning, it is possible to have the pre-trained layers frozen. This has the effect that the optimization process will not alter any of the weights in the pre-trained layers, and only the newly added layers will be updated. Opposite to this is to unfreeze all layers. This results in all parameters in the model being updated during fine-tuning. It is also possible to freeze most of the layers in the model but unfreeze some of the last layers. Freezing the pre-trained layers results in faster training as fewer parameters need to be learned. However, by unfreezing the weights, it may allow better adaptation to Task B, at the cost of longer training time. Depending on the dataset for Task B, unfreezing may also result in overfitting. Another advantage with freezing the pre-trained parameters and a reduction in trainable parameters is that the model consumes less memory on the GPU, which may be limited in some circumstances.

Because training a model from scratch on Task A is very time-consuming and requires advanced and expensive hardware, it is common for researchers to share the pre-trained weights and models. These are referred to as pre-trained models and are openly shared for others to download and leverage in their research. For computer vision and image processing tasks, most of these pre-trained models are pre-trained on the ImageNet dataset.

Chapter 4

Data material

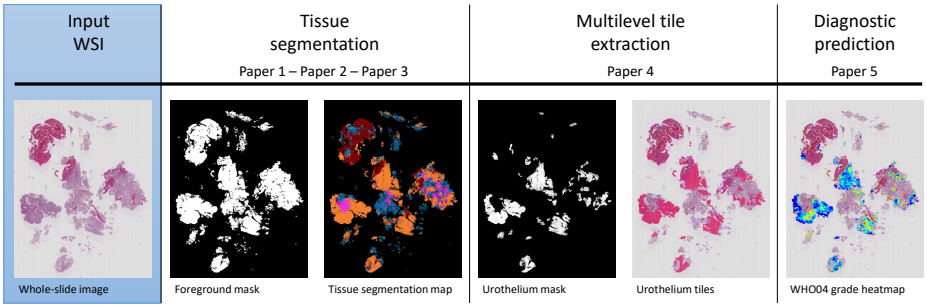


Figure 4.1: An overview of the proposed pipeline, where the topic of this section is highlighted.

This chapter presents the data material used in this work. The histological whole-slide images, highlighted in Figure 4.1, acts as the input to the proposed pipeline. An overview of the process used to create the WSIs is presented, the quality aspect of the material, how the annotation process was conducted, the ethical approval, and finally, an overview of how the data material was split into the different datasets is shown.

In this thesis, the *data material* refers to the 314 WSIs, whereas the different *datasets* refer to subsets extracted from the data material.

4.1 Histological whole-slide images

The data material consists of 314 digital whole-slide images from patients diagnosed with primary papillary NMIBC. All slides are from unique patients, where the tissue is removed through transurethral resection of bladder tumor. The data were collected at the Department of Pathology, Stavanger University Hospital, Norway, in the period between 01.01.2002

and 01.01.2011. The biopsies were formalin-fixed and paraffin-embedded, from which 4 µm thick sections were cut and stained either with hematoxylin, eosin, and saffron (HES) or hematoxylin and eosin (H&E). All WSI have gone through a manual quality check at the department of pathology, Stavanger University Hospital, and only high-quality slides, with little or no blur, have been included in the data material. All WSI are from the same laboratory, and the variation in staining color is relatively low.

All slides were diagnosed and graded according to WHO73 and WHO04 [7], cancer stage, and follow-up data on recurrence and disease progression are recorded. Only patients diagnosed with stage Ta or T1, i.e., non-muscle invasive bladder cancer, are included in the data material. A small section of the available slide-level diagnosis are shown in Figure 4.2.

Year	WHO73	WHO04	Stage	Recurrence	Time to rec	Stage progr	Time to progr
2009	Grade 1	Low grade	TA	no	75	no	79
2010	Grade 1	Low grade	TA	yes	66	no	76
2004	Grade 2	Low grade	TA	no	32	no	49
2007	Grade 2	Low grade	TA	no	74	no	102
2009	Grade 2	Low grade	TA	no	80	no	83
2009	Grade 2	Low grade	TA	yes	8	no	18
2010	Grade 2	Low grade	TA	yes	4	yes	4
2006	Grade 3	High grade	T1	no	109	no	116
2004	Grade 3	High grade	T1	yes	7	yes	59
2009	Grade 3	High grade	T1	yes	8	yes	26
2007	Grade 3	High grade	TA	no	93	no	111
2009	Grade 3	High grade	TA	no	76	no	85
2004	Grade 3	High grade	TA	yes	4	yes	16
2003	Grade 3	High grade	T1	yes	1	no	50

Figure 4.2: A small section of the available diagnostic labels for the data material. The WHO04 labels were used in this work.

4.2 SCN format

The WSIs are captured using a slide scanner with the onboard image processing sensor, which creates a digital image file, much like a regular digital camera. However, the captured images produced by a slide scanner are much larger than a regular photograph and much more complex, with many levels incorporated in the same file. Because of these differences, a slide scanner can not store the captured image in a typical image format like JPEG or PNG. Instead, it is common that the slide scanner stores the images in vendor-specific file formats designed for these images.

Examples of vendor-specific file formats are the slide scanners manufactured by Zeiss, which saves the images using the CZI image format [148], or by Hamamatsu, which stores the captured images in an NDPI file format [50]. There also exist standards for storage and transmission of medical images, like, for example, the Digital Imaging and Communications in Medicine (DICOM) standard [85].

The slides are digitized using a Leica SCN400 slide scanner, and stored in the vendor-specific SCN file format. These images use an XML file to define the structure within the file, which consists of the image pyramid, dimensions of each level, and resolution, to name a few. The images within the SCN file are stored in a single-file pyramidal tiled BigTIFF image. While regular TIFF format uses 32-bit pointers to store the offset values, limiting the file size to 4 GB, the BigTIFF format, however, uses 64-bit offsets and supports file sizes up to 18 exabytes (1.8×10^{10} GB) [11].

These WSI images are gigapixel images with a typical size of $100\,000 \times 100\,000$ pixels, stored as a pyramidal tiled image with several down-sampled versions of the base image in the same file to accommodate for rapid panning and zooming. In the SCN format, each level in the file is down-sampled by a factor of 4 from the previous level. The pyramidal structure of the WSI with three levels is depicted in Figure 4.3.

Due to the vendor-specific file format, WSIs can not be opened using traditional image software. Instead, specific software is needed to read the images. To open the SCN files, Leica has a software called Aperio ImageScope SCN Viewer, which was used as the primary WSI viewer at the beginning of this work [67]. See section 4.5 for more info on the software.

4.3 Magnification and resolution

To examine the stained specimen, it is necessary to enlarge the apparent size of the tissue, allowing a pathologist to view the individual cells of the tissue. This is done either by a microscope or by zooming in a digital WSI to the desired resolution. The ratio between the apparent size of the tissue, and its true size, is called optical magnification. It is a dimensionless number but usually referred to as the power of magnification (e.g., 400x magnification).

For a microscope, a series of lenses and light is used to magnify an object. The eyepiece contains an ocular lens, often with a 10x magnification power. Above the platform is a rotating wheel containing one to five objective lenses

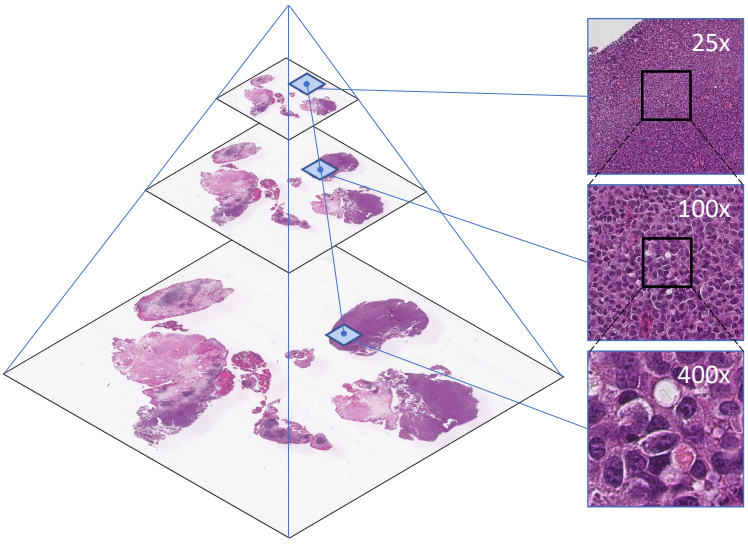


Figure 4.3: WSI images are stored in a pyramidal format, where the base image corresponds to the highest magnification level. The right-hand side shows a set of three tiles extracted so that the center of the tile corresponds to the same physical area in the WSI.

with different magnification powers. The total magnification is the product of the ocular lens magnification and the objective lens magnification. For example, a 10x ocular lens and a 40x objective lens will produce a total magnification of 400x.

Some papers refer only to the objective magnification when describing what resolution they use [19, 78, 115], while others use the total magnifications [33, 74, 146]. All magnification levels in this thesis are referring to the total magnification scale.

Even more confusing is the fact that slide scanners by different manufacturers create WSIs where the stated total magnification does not reflect the apparent size of the produced images. Even though this is not a huge problem in this work, as all WSIs in the data material are produced at the same laboratory, using the same slide scanner. However, it is mentioned because it is important to be aware of the situation, in the case any of the proposed models should be evaluated on external data. To try and avoid confusion, Sellaro et al. [108] suggest instead using microns/pixel as a reference point.

In this work, tiles from different magnification levels are used, depending

on the paper. In Paper 1, only tiles from magnification level 400x are used, while the remaining papers utilize tiles from three levels. The levels in the image pyramid correspond to a magnification level of 25x, 100x, and 400x magnification, which is equivalent to a spatial resolution of 4 µm/pixel, 1 µm/pixel, and 0.25 µm/pixel, respectively.

For the tissue models in Paper 1-3, we used a tile size of 128 × 128 pixels, which for the three magnification levels correspond to (512 µm × 512 µm), (128 µm × 128 µm), and (32 µm × 32 µm). Example tiles can be seen in Figure 2.2.

For the diagnostic model in Paper 5, we had access to a much larger library of WSIs, and thus a larger tile size of 256 × 256 pixels was chosen. For the three magnification levels, this corresponds to (1024 µm × 1024 µm), (256 µm × 256 µm), and (64 µm × 64 µm). Example of such tiles can be seen in Figure 2.3.

4.3.1 VIPS image library

Aperio ImageScope was exclusively used for viewing and annotation of the WSIs. However, by using proprietary software, one is limited to the functionality within the software. It was, therefore, necessary with additional software to be able to process the images in Python. For this, the open-source image processing software called VIPS (VASARI Image Processing System) [84] was used, which has a Python binding called PyVips [101].

The use of the VIPS library allows us to open the SCN images in a Python environment. It has a long list of supported image processing functions and can extract the base image from the SCN-file and the down-sampled versions, a helpful ability used in developing the multiscale methods.

One of the main reasons to use the VIPS library, besides opening the SCN-files, lies in its architecture and multi-threading capabilities. At the core of VIPS is the image-processing library called libvips, which is very memory efficient and fast. Instead of loading the entire image into the computer's memory, it only loads the specific parts of the image which need processing. Also, the library does not execute one command at a time; instead, it stores each command in a pipeline. When the end of the pipeline is connected to a destination, the entire pipeline is executed at once in parallel. In a benchmark comparison, the PyVips library was more than seven times faster than OpenCV [102]. Another comparison from

[101] specifies that "*PyVips is typically 3x faster than ImageMagick and needs 5x less memory*".

4.4 Tissue and image quality

The produced digital slides are not without faults and may include artifacts that impact the overall quality. The quality of a WSI can be divided into two main components; tissue quality and image quality.

The first component, tissue quality, is the quality of the tissue specimen which is placed on the glass slide. Unwanted artifacts may be introduced here either from the TURBT procedure or the preparation of the specimen. Examples of tissue artifacts include cauterized or damaged areas, folded tissue, torn tissue, pen marks, or other artifacts in the WSIs. Inconsistent stain coloring is also a quite common problem. Some examples are shown in Figure 4.4. Because of the cauterization process, WSIs from bladder cancer often contain more damaged areas than other cancer types, like breast cancer or prostate cancer. Also, bladder cancer WSIs include non-diagnostic classes like blood. In addition, muscle tissue and stroma are used for staging of urothelium carcinoma but not for grading. If any of these unwanted tissue classes or foreign objects were included in a diagnosis system, they could negatively impact the result.

The second component is the quality of the image, and artifacts may be introduced during the scanning of the glass slide. Image-based artifacts include areas with blur and out-of-focus areas. To mitigate such inconsistencies, the pathology department, which has produced the WSIs, has strict guidelines to produce WSIs of high quality. WSIs are inspected post scanning, and if there are any problems with the WSI, it will be rescanned. Also, high-quality equipment manufactured by Leica is used for the scanning. We believe the WSIs included in our dataset consist of high image quality; however, small parts may be compromised. An evaluation of the image quality has not been conducted on the dataset used in this manuscript, but this is considered for the future.

Because tissue quality is a much more significant problem in our dataset than image quality, we have focused on developing our tissue segmentation algorithm and then utilizing this for detecting and separating high-quality urothelium tissue from other unwanted areas in the WSI.

The image quality of the dataset used in this study will be evaluated in the future. In some rare cases, issues like out-of-focus or noise can appear

in parts of the WSI. It is both of interest to quantitatively measure the amount of this and see how the proposed system reacts to those areas. A system for identifying areas with image quality could also potentially be used to mask out these areas if they negatively impact the performance of the proposed system.

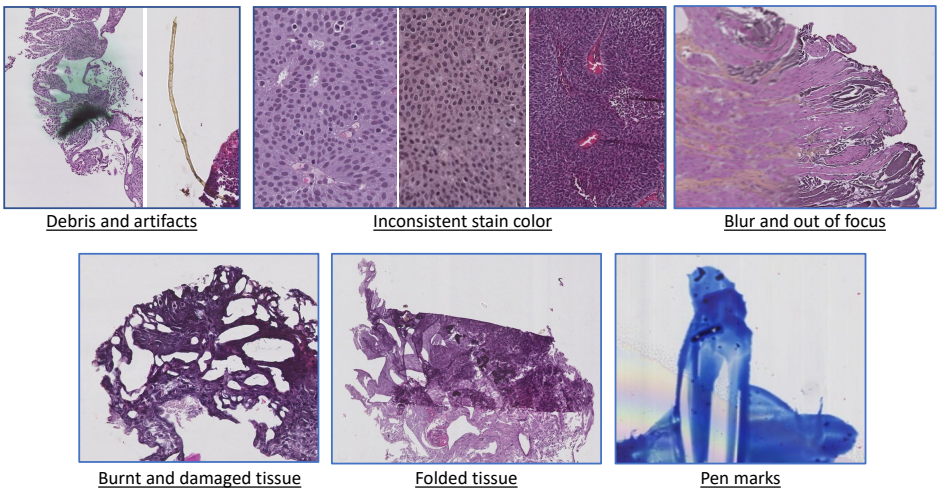


Figure 4.4: Examples of tissue- and image artifacts found in some of the histological images.

4.5 Annotations

Labeled ground truth data is important both for training the models but even more critical for evaluations of the models. The labeling process in this work used two approaches. First, the Aperio ImageScope was used, and later, an in-house developed software called UiS-histology was used.

4.5.1 Aperio ImageScope

Since the WSIs in the data material were scanned using a Leica slide-scanner, it was natural to use Leica's own slide viewer program called Aperio ImageScope SCN Viewer [67]. In addition to navigating the WSIs, the program has a lot of other features. The most helpful feature for us was the free-hand drawing tool. This tool was used to annotate polygon regions in the WSI, and the coordinates were stored in an XML file. A

Python function was created to read the coordinates and create a binary mask of the regions. A screenshot of the ImageScope software with an example annotation region is shown in Figure 4.5.

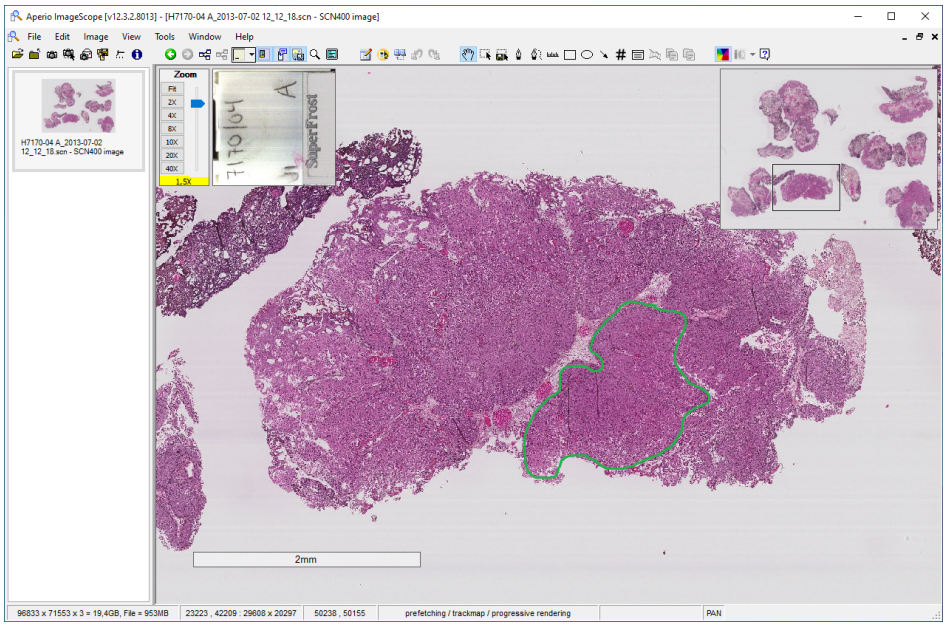


Figure 4.5: The Aperio ImageScope software is used in this work for viewing and annotating whole-slide images. The software has a free-hand drawing tool used for annotating regions on the WSI. A region is shown in green for demonstration (but is not used).

Even though the free-hand drawing tool was working well, the use of proprietary software provided some challenges. First, both the WSIs and software needed to be stored on the same computer. Secondly, because of strict guidelines at the hospital, installation of proprietary software was not possible. These challenges were solved by transferring a subset of the WSIs to a laptop, installing the software, and bringing the laptop to the hospital for annotation.

4.5.2 UiS-histology

Due to the problems mentioned with using a proprietary program, and because other projects at the Biomedical data analysis laboratory (BMDLab) at the university utilize histological images, there was a need for an easier

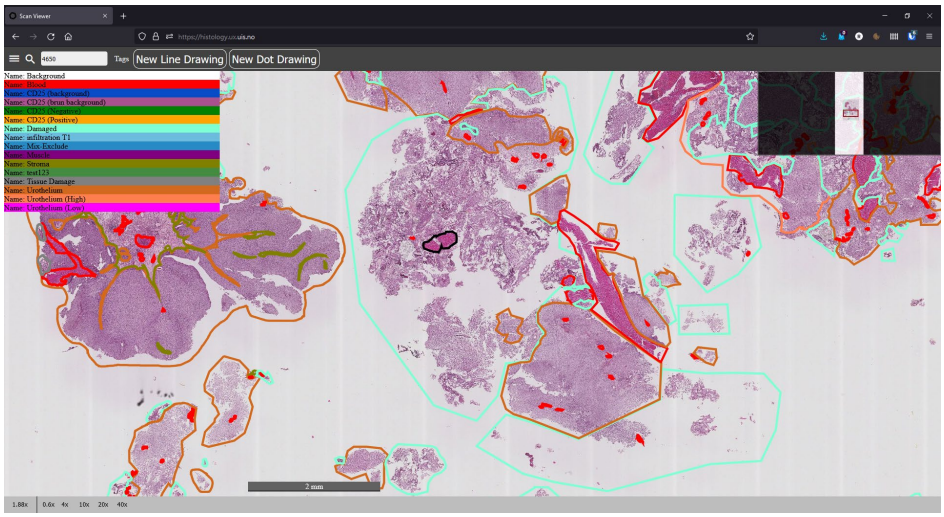


Figure 4.6: The UiS-histology software developed at UiS is accessible via an internet browser. The tools allow for viewing and annotating WSIs, as well as giving specific TAGs to each region. The screenshot shows multiple regions with different types of TAGs illustrated with different colors annotated by a pathologist.

way to handle this. An in-house tool for viewing and annotating WSIs was developed at the department, and about halfway through this thesis work, a new annotation tool was available. This tool is referred to as UiS-histology, and a screenshot can be seen in Figure 4.6. The UiS-histology tool has many advantages. All WSIs are stored locally at the university and are remotely available through a web browser. In addition, free-hand annotation is possible without installing any software, and annotated regions can be tagged with different classes (e.g., "urothelium," "muscle tissue," or "high-grade").

4.6 Ethical approval

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. As this is a retrospective study, Ethical approval was given without written consent from the patients. All insights in a patient's journal are monitored electronically, and all except the treating physician were required to state the reason why they needed to read that patient's journal. This log is always open for the patient to view.

All patients were checked if any had registered themselves in the register for research reservation from the National Institute of Health (Registry of Withdrawal from Biological Research Consent, Norway).

4.7 Dataset overview

The data material described in this chapter, consisting of 314 WSIs and diagnostic labels, is the basis for all the datasets used in the published papers. An overview of how the WSIs are distributed is shown in Figure 4.7, and a description of each dataset follows below. A more detailed description of how tiles are extracted is given under each respective paper in sections 5, 6, and 7.

The names Dataset A, B, C, . . . , are not used in the papers but introduced here for improved readability and will be used throughout the thesis.

For training and evaluation of the various models, several training, validation, and test sets have been constructed. Some of the WSIs are present in more than one dataset; however, care has been taken into account to ensure no cross-contamination between any of the training and testing sets of the same model.

WSIs included in Dataset E, F, and G were randomly selected and stratified to include the same ratio of all diagnostic outcomes based on the WHO73 and WHO04 grading, stage, recurrence, and disease progression, to represent the data material best.

Dataset A – Training

In Dataset A, 48 WSIs were extracted randomly from the data material and used as unlabelled training data to train an autoencoder in Paper 1.

From the 48 WSIs in Dataset A, 26 are also present in Dataset B, two in Dataset C, three in Dataset D, 35 in Dataset E, seven in Dataset F, and four in Dataset G. The WSIs shared between Dataset A and B, are only used for training data in Paper 1, and separate WSIs are used for testing. The WSIs shared between Dataset A and the remaining datasets C-G are separate models.

Dataset B – Training, validation and test

Dataset B consists of 32 WSIs with a total of 239 annotated regions belonging to the five foreground tissue classes (urothelium, stroma, muscle, damaged tissue, and blood). The dataset was created using the Aperio ImageScope over multiple sessions at Stavanger University Hospital. Together with a pathologist, we annotated regions in as many WSIs as possible over the limited time scheduled for the task. Regions for the background class were randomly selected afterward and are different between the papers.

Tiles are extracted from the annotated regions and act as the basis dataset used in Paper 1-3. For Paper 1 and 3, the data is split on WSI level into training and test set, and for Paper 2, all data is used as training and validation data using cross-validation. Because the three papers utilize different methods (autoencoder, cross-validation, and semi-supervised learning, respectively), the number of extracted tiles differs between the papers.

Some tissue classes are more sparse in the tissue specimen, and thus harder to find large regions suitable for annotation. This creates a class imbalance, where the classes of stroma and muscle tissue have fewer samples than the remaining classes.

The dataset was annotated on the 400x magnification scale and is considered strongly labeled on this level. In Papers 2 and 3, the dataset is used with multiscale models utilizing lower magnification levels (25x, 100x), and by keeping the tile size the same, the lower magnification tiles will have a wider field of view, allowing for more context of the surrounding tissue to be included. Consequently, these tiles will, in some cases, include several classes. For these scales, the dataset is considered weakly labeled.

From the 32 WSIs in Dataset B, 29 WSIs are present in Dataset E, and three are present in Dataset F. Dataset E and F are the training and validation datasets for the diagnostic model, where the TRI-model trained on dataset B is used as the tissue model. However, none of the WSIs present in Dataset B is present in the test dataset for the diagnostic model (Dataset G).

Dataset C – Test

This dataset consists of seven unlabeled WSIs, used for testing the tissue model in Paper 2. From the seven WSIs in Dataset C, it shares two WSIs

with Dataset A and two WSIs with Dataset D. However, these datasets are used on separate models.

Ideally, we would have annotations of all the tissue classes in these WSIs, but this was not possible. Later, after Paper 1-3 were published, using the UiS-histology annotation tool, a pathologist annotated all tissue in one example WSI from Dataset C.

Dataset D – Training

This dataset consists of 46 unlabeled WSIs, used for training data in Paper 3. From the 46 WSIs in Dataset D, three WSIs are present in Dataset A, two in Dataset C, 29 in Dataset E, two in Dataset F, and five WSIs in Dataset G. However, these are used to train separate models, not in conflict with each other.

Dataset E – Training

This dataset consists of 220 WSIs and corresponding WHO04 labels used to train the diagnostic model in Paper 5. It consists of 124 low-grade and 96 high-grade WSIs.

From the 220 WSIs in Dataset E, 62 WSIs are present in either Dataset A or D; however, these are unrelated to each other. In addition, 29 of the WSIs in Dataset E are present in Dataset B. Dataset B is used to train the tissue model used in the diagnostic pipeline.

Dataset F – Validation

This dataset consists of 30 WSIs and corresponding WHO04 labels used as validation data for the diagnostic model in Paper 5. It consists of 17 low-grade and 13 high-grade WSIs.

From the 30 WSIs in Dataset F, nine WSIs are present in either Dataset A or D; however, these are unrelated to each other. In addition, two of the WSIs in Dataset F are present in Dataset B. Dataset B is used to train the tissue model used in the diagnostic pipeline.

Dataset G – Test

This dataset consists of 50 WSIs and corresponding WHO04 labels used as test data for the diagnostic model in Paper 5. It consists of 28 low-grade and 22 high-grade WSIs.

In addition to the slide-level WHO04 labels, a pathologist has annotated 30 low- and high-grade regions in 14 WSIs (seven low-grade and seven high-grade WSIs), in a subset of Dataset G, referred to as the segmentation test set.

From the 50 WSIs in Dataset G, four of the WSIs are present in Dataset A, and five in Dataset D. These datasets are unrelated.

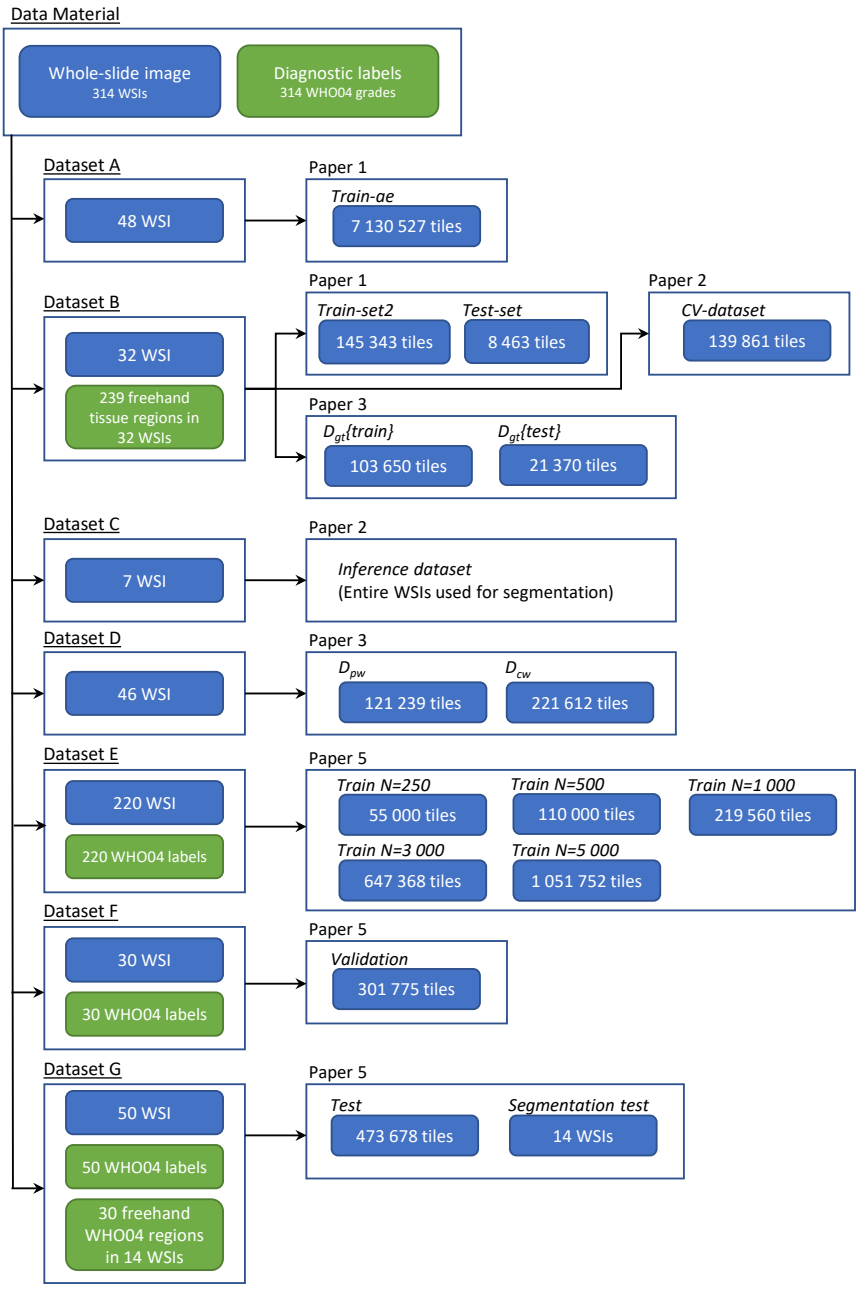


Figure 4.7: Overview of all datasets used in each paper. Which paper each dataset is used in is written on top of the box, and the name of the dataset in the paper is specified inside the box.

Chapter 5

Tissue segmentation

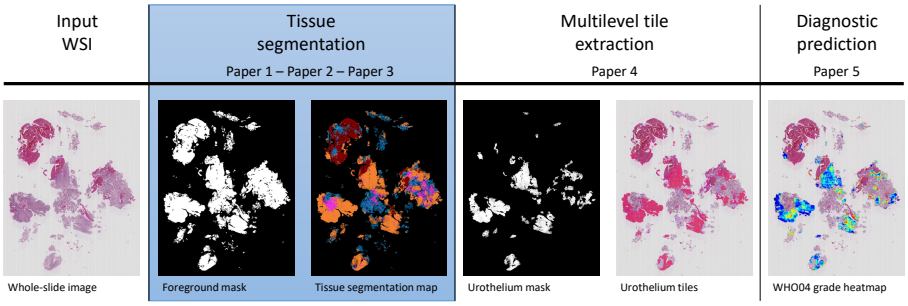


Figure 5.1: An overview of the proposed pipeline, where the topic of this section is highlighted.

The following chapter is dedicated to the topic of tissue segmentation; the blue highlighted section in Figure 5.1. Paper 1-3 are all dedicated to the topic, and the main methods, results, and contributions from each will be presented. In the thesis, these papers are part of the sub-objectives SO_1, SO_2 and SO_3.

SO_1: Create an automated system for distinguishing between the different tissue types present in histological whole-slide images of urothelial carcinoma.

SO_2: Explore different approaches for unsupervised and semi-supervised learning techniques to deal with the lack of detailed region-based annotation data.

SO_3: Investigate the use of multiscale models in WSI processing by utilizing several magnification scales.

5.1 Contribution overview

The objective is to differentiate the tissue classes in WSI from NMIBC automatically. A set of six classes were selected in cooperation with pathologists at the Stavanger University Hospital: urothelium, stroma, muscle, damaged tissue, blood, and background. The urothelium is the diagnostic relevant tissue class for grading bladder cancer, and urothelium, stroma, and muscle tissue are all used in staging bladder cancer.

A system capable of identifying these tissue types and visualizing their location in the WSI will have multiple benefits. First, it can guide the pathologists to the diagnostic relevant areas of the WSI, making their workflow more efficient. Also, it can be used to find and automatically extract these diagnostic relevant areas and used as input in a computer-aided diagnostic (CAD) system. Damaged tissue can potentially impact the diagnostic predictions negatively, and a system should therefore be able to identify these areas and exclude them from further analyses. Furthermore, identification of muscle tissue in the WSIs would also be beneficial. It plays a vital role in the staging of bladder cancer, as pathologists want to know whether the tumor has infiltrated the muscle tissue. Also, in the pathologic report, pathologists must specify whether muscle tissue is present or absent in the specimen [6]. Muscle tissue is usually relatively sparse in the WSI, and it can be time-consuming to get a complete overview of its locations. However, with the help of segmented tissue images, it can be verified in a short amount of time.

5.2 Paper 1 – Autoencoder

In this section, the contributions in Paper 1 are presented, where a method for automatic classification of WSI into six different classes is proposed. The method is based on CNN, firstly trained unsupervised using a large unlabelled dataset by utilizing an autoencoder. Thereafter, a smaller labeled dataset is used to fine-tune the final fully-connected layers from the low dimensional latent vector of the autoencoder, providing an output as a probability score for each of the six classes, suitable for automatically defining regions of interests in WSI. The principle of autoencoder models is explained in Chapter 3.2.2.

5.2.1 Data material

The proposed method in Paper 1 uses two datasets, referred to as Dataset A and Dataset B in Figure 4.7. For both datasets, non-overlapping tiles of size 128 × 128 pixels are extracted at the 400x magnification level, corresponding to a spatial resolution of 32 µm × 32 µm.

The first dataset is a large unlabeled dataset used to train the autoencoder model from scratch. This dataset is referred to as *train-ae*, and consists of 7 130 527 unlabeled tiles extracted from the 48 WSIs in Dataset A. On the other hand, Dataset B is a strongly labeled dataset used to fine-tune the encoder-classifier model. To compensate for the class imbalance, tiles belonging to the stroma and muscle classes were augmented by using overlapping tiles during extraction and rotation and mirroring the extracted tiles. Augmentation was not performed on the test set.

Dataset B was split into a training set with 145 343 tiles after augmentation referred to as *train-set2* and a test set with 8 463 tiles called *test-set*. The split was done on WSI-level, and none of the WSIs from Dataset A or *train-set2* was part of the *test-set* to avoid cross-contamination between training and test data.

5.2.2 Method

The architecture of the best performing encoder-decoder model is depicted in Figure 5.2. The encoder consists of several convolutional layers, max-pooling layers, dropout, and fully-connected layers. The decoder consists of the same layers but in reverse order and uses unpooling and deconvolutional layers instead. An extensive grid search was conducted as the first experiment in the paper to find the optimal number of convolutional layers and the size of the latent vector. In Figure 5.2, a tile is extracted from the input WSI and used as input to the autoencoder. On the right-hand side, the reconstructed output tile is visible.

There is a strong correlation between the size of the latent vector and the loss between the input and reconstructed output. With a large enough latent vector, the loss approaches zero, and almost perfect reconstruction can be achieved. However, the goal is not to use the latent vector for reconstruction but rather for classification in the next step. Hence, a small latent space was chosen, which force the network to extract only the essential features of the input and preserve these in the vector.

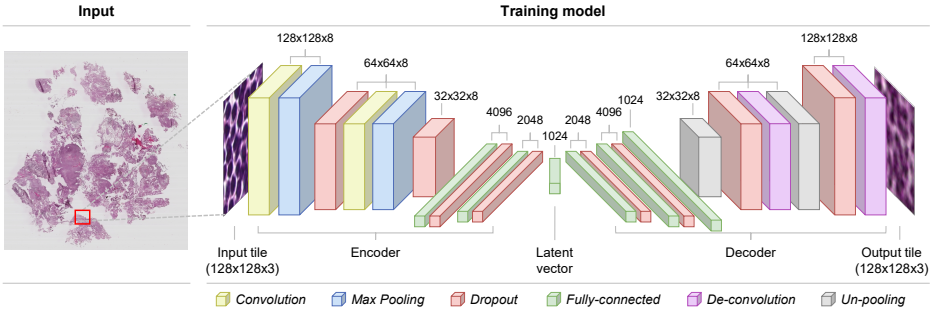


Figure 5.2: The autoencoder model consists of two main parts: the encoder and the decoder. Tiles are extracted from the WSI and fed to the encoder network, compressing the data into the latent vector. The decoder network will reconstruct the input image from the latent vector. The difference between the input and reconstructed output constitutes the loss of the model used to train the model. This step is referred to as pre-training of the model.

After training the encoder-decoder network, the decoder part is substituted by a new classification network as depicted in Figure 5.3. The classification network consists of three fully-connected layers, where the output layer uses a softmax activation function, yielding a probability score for each class. This encoder-classifier model constitutes the proposed CNN-model.

To find a suitable architecture and appropriate hyperparameters, a large grid search was conducted consisting of 36 different encoder-decoder models and 162 encoder-classifier models. The grid search included different latent vector sizes, learning rates, dropout rates, different numbers of convolutional layers in the encoder and decoder, different numbers of fully-connected layers in the classifier, and freezing and unfreezing the encoder during fine-tuning. The specific values for all parameters are reported in Paper 1.

To reduce both computational time and search space, a preliminary search was set up with some limitations. A reduced version of the *train-ae* dataset was used to decrease the processing time, and each model was only trained for 50 epochs, and hyperparameters that showed poor performance on several models were excluded to narrow down the search space.

The resulting architecture from the grid search was trained once more. First, the autoencoder was trained on the full *train-ae* dataset for 100 epochs, followed by fine-tuning of the encoder-classifier on the *train-set2* dataset for another 600 epochs. Since the grid search showed the best results

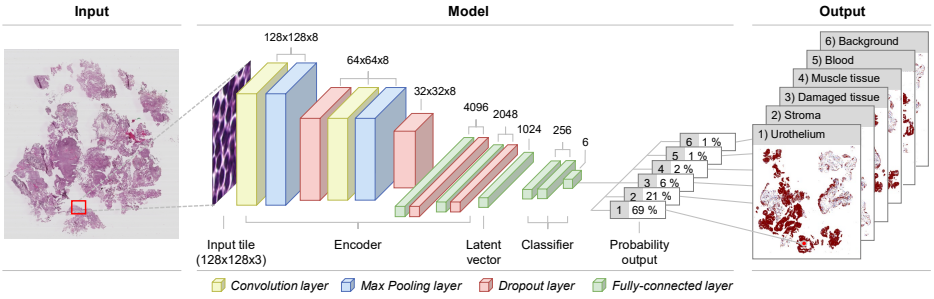


Figure 5.3: The decoder network is exchanged with a classification network. The encoder network will compress the input tile to the latent vector, and the classification network will classify the tile into one of the six classes. During training, this step is referred to as fine-tuning. The finished model can classify new WSIs by outputting a probability score for each of the six classes used to create heatmaps.

when the encoder was not frozen during fine-tuning, both the encoder and classifier were trained during this step.

The final model can classify entire WSIs tile by tile, with or without overlapping, and produce heatmaps, visualizing each tissue class and its location in the image. Such maps can provide useful information to a pathologist during visual inspection. As seen in Figure 5.3, one heatmap is created for each class. The maps were first filtered using a gaussian filter, and then a thresholding operation was performed with a limit of 0.8, setting all predictions below this threshold to zero. This ensures that only predictions of 0.8 or higher are visible in the final heatmaps.

5.2.3 Result

The proposed CNN-model achieved an F1-score of 93.8% on the urothelium class and an average F1-score of 93.4% over all six tissue classes. The precision, recall, and F1-score of each class is shown in Table 5.1.

The model was further used to create heatmaps from three unseen WSIs, and visualize each tissue class in the image. Figure 5.4 shows the WSIs with their corresponding heatmaps. Only the urothelium heatmaps are shown here, but heatmaps for all classes are shown in Paper 1.

Table 5.1: Detailed classification results from the model trained using 10% dropout rate.

Class	Precision	Recall	F1-Score
Urothelium	0.924	0.952	0.938
Stroma	0.897	0.929	0.913
Damaged	0.925	0.927	0.926
Muscle	0.980	0.714	0.826
Blood	0.996	0.991	0.994
Background	0.990	0.988	0.989
Average total	0.936	0.935	0.934

5.2.4 Conclusion

This paper proposes a method for automatically classifying tile-segments of histopathological WSI of urinary bladder cancer into six different classes using a CNN-based model.

The use of autoencoder for pretraining followed by supervised fine-tuning for tissue classification seems promising with an F1-score of 93.4%. Performance of the heatmaps can not be quantitative measured but have been visually inspected by pathologists and is considered very promising.

Further work will include an effort to improve the classifier, and other methods such as a multiscale approach are considered.

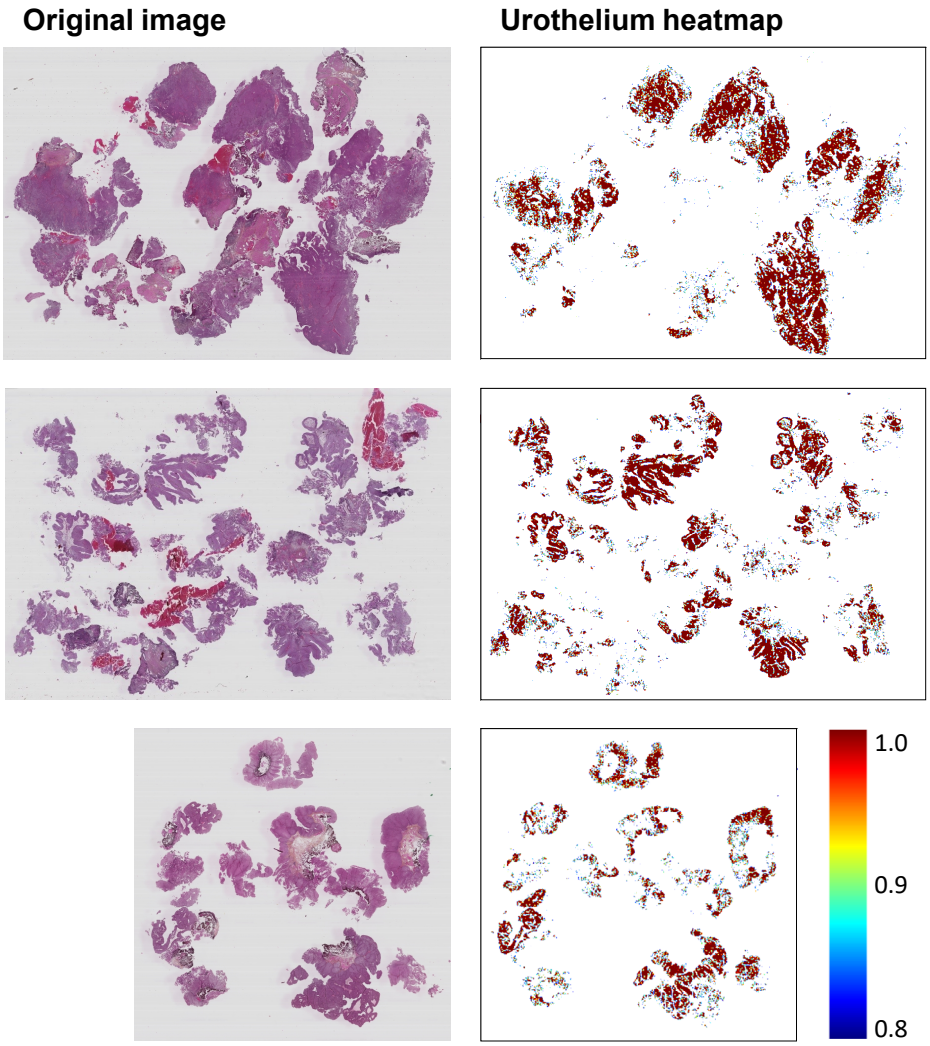


Figure 5.4: Urothelium heatmaps shown for three WSIs. Heatmaps for the remaining tissue classes are shown in Paper 1. Note that only urothelium with a probability score of 80% and higher are shown.

5.3 Paper 2 – Multiscale model

In this section, the contributions in Paper 2 are presented. The objective of this paper was to explore if it was possible to reduce the training time by leveraging on pre-trained models rather than training an autoencoder model

from scratch. A second objective was to investigate the use of multiscale in the input tiles. The tile size is kept constant over the scales so that the field of view is increased with decreased resolution.

Tiles are extracted from three magnification scales: 25x, 100x, and 400x. When a tile is saved from the region, the corresponding tiles from 25x and 100x magnification were also extracted in such a manner that the center pixel is the same in all three magnification levels, as can be seen in the right-half of figure 4.3.

5.3.1 Data material

The methods in this paper are trained and evaluated on Dataset B from Figure 4.7, and further demonstrated on Dataset C.

Due to the small size of Dataset B, stratified 5-fold cross-validation was implemented. This way, all tiles are used both for training and validation. Stratification is performed on slide-level to ensure that tiles from the same patient are not present in both the training and test set. A fixed seed is set to ensure that the folds are the same for each model, making the included samples in the training and test sets identical for all models. Background regions were annotated in five additional WSIs, to fit the 5-fold cross-validation scheme better. Tiles belonging to the stroma- and muscle-tissue classes were augmented similarly to the description in Paper 1 but with less rotation and mirroring.

A new strategy for extracting tiles from all classes was implemented. This strategy, referred to as automatic grid search, was implemented to better accommodate the shape of the tissue regions in the WSI, as these often contain tight corners and narrow passages. The grid of non-overlapping tiles was shifted in the X- and Y-direction to maximize the number of extracted tiles. This search was performed individually for each of the 239 regions.

Stratification means that the number of WSIs for each class is evenly distributed across all five folds. However, some of the WSIs contained annotation for several classes in the same WSI, and by evenly distributing the WSIs containing the urothelium regions, it may result in uneven distribution of another class. To achieve the best possible stratification of this dataset, an automatic python-script was implemented that tested all combinations by brute force. Once the optimal WSI distribution was found, the seed creating the split was saved and used throughout the experiments.

By collecting tiles from multiple scales, the model has access to more context information from the surrounding tissue. However, by keeping the tile size the same, the lower magnification (25x, 100x) tiles will have a wider field of view, allowing for more context of the surrounding tissue to be included. Consequently, these tiles will, in some cases, include several classes. Furthermore, the labels are imprecise as they do not include samples of the labeled border between tissue regions. This would require multi-label samples, an even more expensive annotation process. As a result of this, the dataset is weakly labeled in both quantity and quality.

With the additional WSIs for background, implementation of the automatic grid search, but fewer tiles gained from augmentation, a total of 139 861 tiles were extracted from the WSIs in Dataset B. In addition, the seven WSIs from Dataset C were used as inference on the re-trained models. The WSIs included in the inference dataset are not part of the CV dataset and thus unseen by the models.

5.3.2 Method

An overview of the proposed system is presented in Figure 5.5 and consists of three steps. First, a binary background mask is produced from the 25x level of the WSI, generated by checking the pixel intensity value and splitting them into background or non-background tiles. The non-background tiles are then extracted and fed to the multiscale tissue model in step 2 for further classification.

To increase the resolution of the resulting segmentation image, the 128 $\times$ 128 pixels tiles are extracted using overlap. The stride of the overlap affects the resolution, and a typical value of 8 pixels is used, but it can be any value within the size of the tile. Tiles are classified according to the highest prediction score. However, the score must be above a specific threshold value to be considered valid. A threshold value of 0.6 was determined as a trade-off between removing unwanted predictions and not removing too much. Tiles with all prediction scores below the threshold are labeled as *undefined*.

Finally, in the last step, after each tile was classified, a color-mapping function was used to give each class a separate color. Then, by combining all the predicted tiles, a segmentation image is created, visualizing the location of all classes.

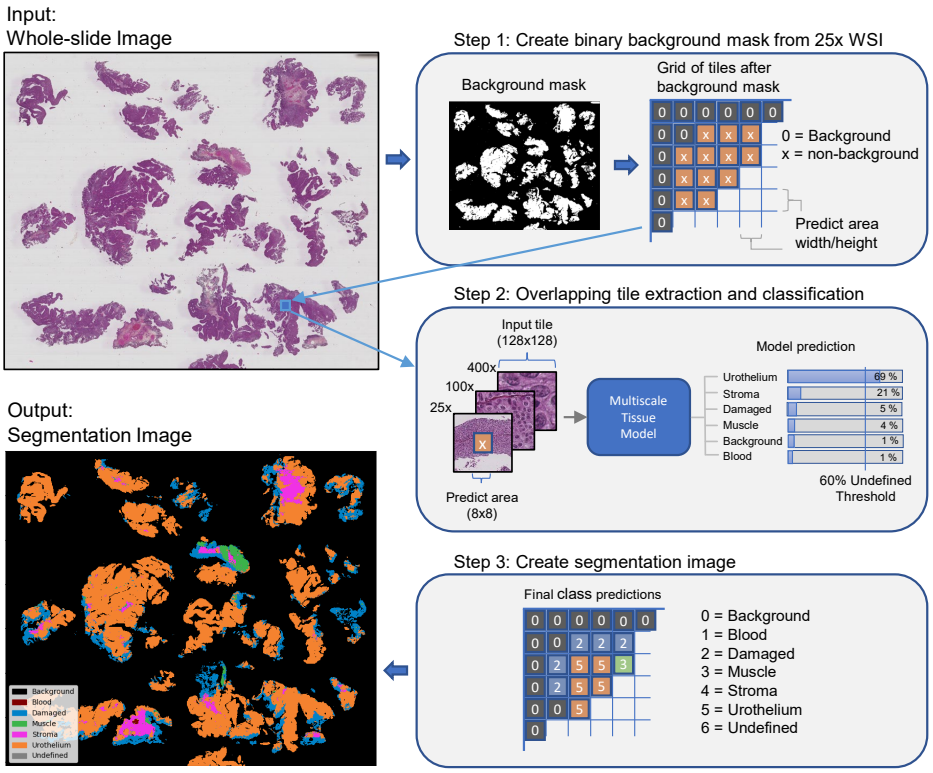


Figure 5.5: Overview of the proposed system in Paper 2. A background mask is created from the 25x WSI to exclude the background from further processing. Areas in the WSI selected as non-background is then extracted and fed through the multiscale model from Figure 5.6, which outputs tissue predictions. The prediction needs to exceed a set threshold to be valid. Finally, the segmentation image is generated by giving each class a separate color. The values shown in the figure are for illustration purposes only.

Model architecture

This paper proposes three architectures referred to as the MONO-, DI-, and TRI-CNN models. The three architectures have one, two, and three inputs, respectively. To differentiate the models from each other, they are named according to their main architecture and the input scale, e.g., MONO-400x is a MONO-CNN model trained on tiles extracted at 400x magnification.

Tiles in the dataset are extracted at three magnification levels, yielding three MONO models: MONO-25x, MONO-100x, and MONO-400x. These three magnification scales can further be combined in three configurations for the DI-CNN model: DI-25x-100x, DI-25x-400x, and DI-100x-400x. The

TRI-CNN model has only one configuration: TRI-25x-100x-400x, and is depicted in Figure 5.6. The different MONO- and DI-CNN models can easily be derived from the same figure. E.g., to create the DI-25x-400x model, remove the 100x input and blue blocks, and to create the MONO-100x model, remove the 25x input, 400x input, red and yellow blocks.

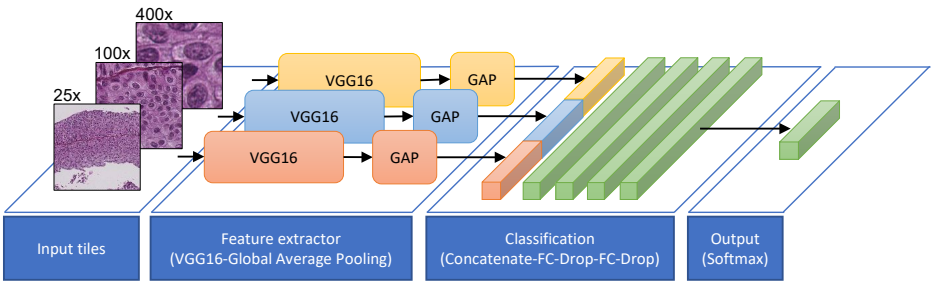


Figure 5.6: A block diagram of the TRI-CNN model proposed in the current paper. The input tiles are fed through individual pre-trained VGG16 network and global average pooling (GAP) layer to create feature vectors. The feature vectors are concatenated and fed through the classification network before entering the final output layer consisting of a softmax function. The softmax function outputs a prediction score for each of the six classes.

The overall structure of each model is the same. Each input is fixed at $128 \times 128 \times 3$ pixels, which is the size of each tile. The input is fed into a pre-trained VGG16 network [113] which acts as a feature extractor, followed by a global average pooling (GAP) layer providing a feature vector representation of the input. This feature vector is then fed into a classification network consisting of two fully-connected (FC) layers, each followed by a dropout layer, and a final softmax layer with one output node for each class. The DI- and TRI-CNN models have two and three parallel VGG16 branches, resulting in multiple feature vectors. These feature vectors are concatenated before entering the classification network. The FC-layers have the same size of 4 096 neurons as the original layers in the VGG16 network. Dropout layers are added after each FC-layers to add regularization to the network due to the small dataset.

Experiments

A wide range of experiments was performed. Each magnification scale was tested individually, in addition to all combinations of the three scales. Freezing and unfreezing the parameters of the VGG16 base model were

Table 5.2: Results for all 28 models, trained using stratified 5-fold cross-validation. Each score is shown as micro-averaged F1-score aggregated across all classes, marked as 'All' in the table. F1-score only for the urothelium class is shown in the columns marked 'Uro.'. Numbers in bold refer to the highest score in their respective column.

	Multiclass				Binary-class			
	Frozen		Unfrozen		Frozen		Unfrozen	
Model	All	Uro.	All	Uro.	All	Uro.	All	Uro.
MONO-25x	93.4	92.9	**96.4**	96.8	96.3	92.5	98.1	96.1
MONO-100x	94.4	96.6	94.8	97.8	98.3	96.5	99.1	98.1
MONO-400x	87.2	89.7	86.4	86.3	94.2	88.1	93.7	87.2
DI-25x-100x	**96.5**	97.4	96.2	98.1	98.1	96.2	**99.3**	**98.5**
DI-25x-400x	95.6	96.3	96.0	97.6	97.8	95.4	98.3	96.5
DI-100x-400x	95.0	96.8	95.3	97.6	98.4	96.6	98.9	97.7
TRI-25x-100x-400x	**96.5**	**97.6**	**96.4**	**98.3**	**98.5**	**97.0**	99.2	98.3

tested to see if unfreezing the weights would lead to better adaption to the histological domain at the cost of longer training time. Instead of classifying tiles into six classes, a possible easier problem would be to only classify urothelium vs. non-urothelium tissue. Therefore, each model was also tested with this binary-class approach to see if it improved classification results for urothelium tissue.

After evaluating the model using stratified cross-validation, a new and final inference model was trained by utilizing all available data as training data. The average number of epochs used during cross-validation was used when training the inference model. This inference model was then used to predict new WSIs from the inference dataset.

5.3.3 Result

Table 5.2 shows the cross-validation results for all the models. Aggregated micro-average F1-score across all classes are included, as well as the F1-score for only the urothelium class to better compare multiclass vs. binary-class models.

Based on the results in Table 5.2, some of the best models were chosen to segment the WSIs in the inference dataset. The selected models were

retrained on the entire CV dataset before segmenting the WSIs. Figure 5.7 shows a comparison between segmentation images generated by the best binary-class model and the best multiclass model. A DICE-score is calculated to measure the similarity between the predicted urothelium tissue between these two models, with an average DICE-score of 0.87 for the three WSIs. Segmentation images for the remaining four WSIs in the inference dataset can be viewed in Paper 2.

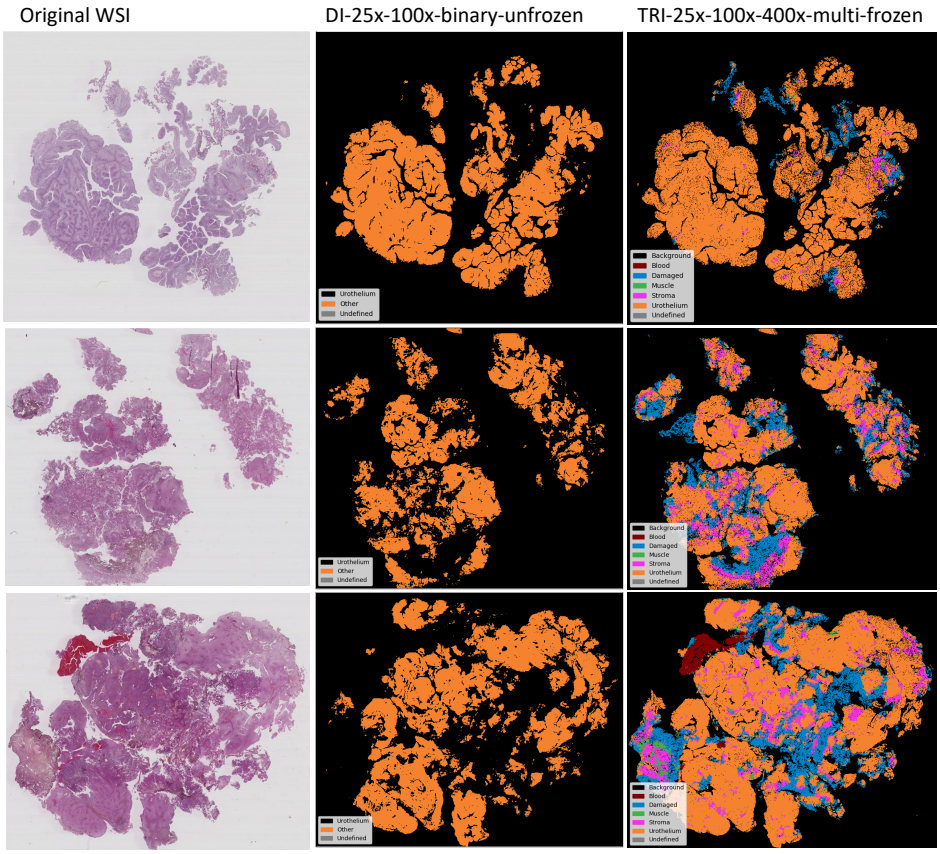


Figure 5.7: The best binary-class model vs. the best multiclass model. A DICE-score is calculated to measure the similarity between the predicted urothelium tissue between the two models. DICE-score from top to bottom are 0.92, 0.85 and 0.85.

5.3.4 Conclusion

This work uses convolutional neural networks (CNN) for multiscale tile-wise classification and coarse segmentation, including both context and detail, by using three magnification levels: 25x, 100x, and 400x. Twenty-eight models were trained on weakly labeled data from 32 WSIs, where the best model got an F1-score of 96.5% across six classes. The multiscale models were consistently better than the single-scale models, demonstrating the benefit of combining multiple scales.

The best models were retrained on all available data and used to segment seven unseen WSIs. This is potentially useful to both help pathologists in their work with diagnosing bladder cancer patients, as well as extracting diagnostic relevant areas of the WSIs of high quality to be used in an automatic computer-aided diagnostic (CAD) system. The resulting segmented images have been manually inspected by an expert uropathologist and are considered to be very promising especially considering that the WSIs were only weakly annotated.

5.4 Paper 3 – Semi-supervised learning

In this section, the contributions in Paper 3 are presented. This paper uses a fixed model architecture and focus on experimenting with different datasets used to train the models. The best architecture from Paper 2, TRI-25x-100x-400x, is used in all experiments.

The paper deals with semi-supervised learning on the application of tissue-type classification by using the model itself to find and extract a larger dataset, which is subsequently used to fine-tune the model. Two semi-supervised approaches utilizing the unlabeled data in combination with a small set of labeled data are presented. The first method is a probability-based method based on predicted probabilities from an initial model. The second method is a cluster-based self-training method based on both predicted probability from the initial model and local neighborhood in the predictions.

5.4.1 Data material

Dataset B from Figure 4.7 is used as a ground truth dataset, and divided into a training and test set. The training set, named $D_{gt}\{train\}$, consists

of 103 650 tiles from 29 WSIs, and the test set, $D_{gt}\{test\}$, consisting of 21 370 tiles from 8 WSIs.

In addition, the 46 unlabeled WSIs in Dataset D were used with the two self-training methods. For the probability-based method, a total of 121 239 tiles were extracted from all 46 patients and formed the probability-weak dataset, D_{pw}. For the cluster-based method, a total of 221 612 tiles were collected from 44 patients and formed the cluster-weak dataset, D_{cw}.

5.4.2 Method

Six multiscale models are presented, and the following letters are used to describe them: SL is short for supervised learning, and SSL for semi-supervised learning. P indicates that the models are trained through the probability-based self-training method, and C implies that the cluster-based self-training method is used. A refers to that augmentation by rotation of tiles is involved. Finally, F and U refer to the weights in the VGG16 models being frozen or unfrozen during training, respectively.

All models are evaluated on the $D_{gt}\{test\}$ dataset. In addition, they are used to segment a new WSI to investigate the model's performance with regard to segmentation. The WSI has been annotated by a pathologist and has not been used during training. This WSI is referred to as WSI_segment_test, and the predictions of the WSI is visually compared to the ground truth annotations.

Two versions of the TRI model were trained through supervised learning on the $D_{gt}\{train\}$ dataset, one with frozen weights and one with unfrozen. These models are referred to as TRI-SL-AF and TRI-SL-AU. The trained models were then evaluated on the $D_{gt}\{test\}$ dataset and acts as a baseline for the two semi-supervised methods. The best model of these two, the TRI-SL-AF, is further used in combination with the SSL methods as described below.

Probability-based self-training

The first SSL method is a probability-based self-training approach, depicted in Figure 5.8. Two models are trained, referred to as TRI-P-SSL-F and TRI-P-SSL-AU.

First, the TRI-SL-AF model is used to classify all tiles from the 46 WSIs in Dataset D. Only tiles with a probability of 60% and higher are saved.

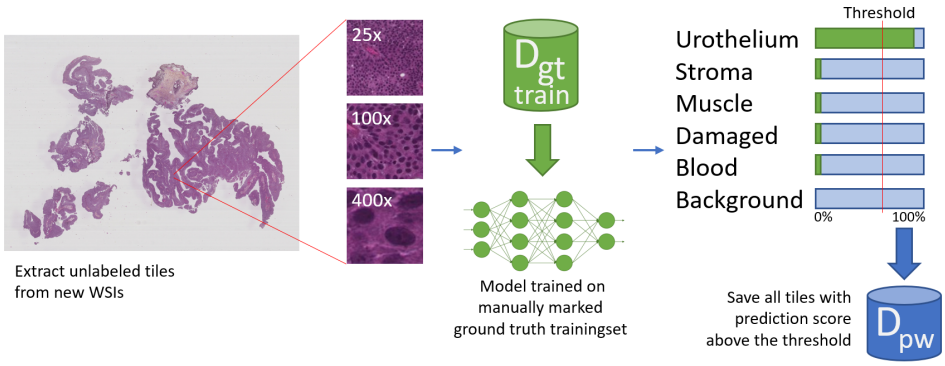


Figure 5.8: Origin of probability-weak dataset, D_{pw}.

Next, a subset of the saved tiles is selected based on four criteria: minimum and maximum tiles per WSI, minimum tile probability, and maximum tiles in total. The criteria value is set for each tissue class, and a complete overview is given in table 2 in the paper. For each patient, tiles with the highest probability are collected first until the maximum number of tiles per WSI has been collected, or no more sufficient tiles remain. The subset of tiles that meets the criteria are saved in the probability-weak dataset D_{pw}.

The two probability-based models, TRI-P-SSL-F and TRI-P-SSL-AU, were trained on the labels in both the $D_{gt}\{train\}$ and D_{pw} datasets.

Cluster-based self-training

The second SSL method is a cluster-based self-training approach, depicted in Figure 5.9. Two models are trained, referred to as TRI-C-SSL-F and TRI-C-SSL-AU.

Similar to the probability-based method, the cluster-based method uses model TRI-SL-AF to classify the WSIs. The tiles are classified with a minimum of 60% probability, and tiles with a lower probability are discarded.

For the cluster-weak dataset, six criteria are given for a tile to be valid. These are the minimum and maximum tiles per WSI, maximum clusters per WSI, minimum cluster size, maximum tiles per cluster, and minimum average cluster probability. A complete overview of the criteria for each tissue class is given in table 3 in the paper.

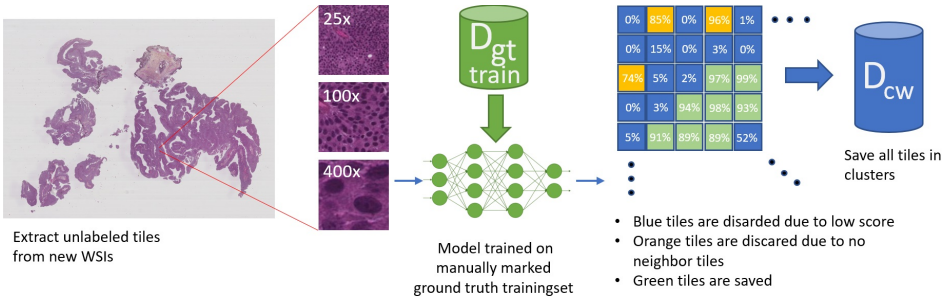


Figure 5.9: Origin of cluster-weak dataset, D_{cw}.

An algorithm searches through the tiles and groups them into clusters. If, at any point in the search, the maximum number of tiles per cluster is not reached, the difference is appended to the limit of the next cluster in line. The average cluster probability is calculated per cluster, and the clusters are sorted after the highest probability. Each cluster originating in the WSI is then sorted into an array, and the program selects the clusters based on the highest probability according to the maximum number of clusters. The labels are then saved to the cluster-weak dataset D_{cw}.

The two models, TRI-C-SSL and TRI-C-SSL-AU, were then trained on labels from both the $D_{gt}\{train\}$ and D_{cw} datasets.

5.4.3 Result

All six multiscale models were tested on dataset $D_{gt}\{test\}$, yielding the results in Table 5.3. The models were also used to segment a new WSI, and close-up regions were compared with the annotated ground truth regions. These segmentation images can be seen in figures 6 and 7 in the paper.

The supervised model, TRI-SL-AF, trained only on the $D_{gt}\{train\}$ dataset, achieved an accuracy of 94.61%. The best SSL model, TRI-C-SSL-AU, trained on both the $D_{gt}\{train\}$ and D_{cw} datasets, improved the accuracy by 1.38% and got a score of 95.99%. Furthermore, the F1-Score stayed the same or increased for every single class, and a distinct improvement is seen in the predicted segmentation maps.

The probability-based model TRI-P-SSL-AU showed an improvement in classifying urothelium, with an increase of 1.44% in F1-Score, from an initial 98.08%. However, the accuracy was only increased by 0.24%, as the model had a reduction in the F1-Score for blood.

Table 5.3: F1-Scores for each of the classes, and overall accuracy for the six models. Green cells indicate the best result within each category. The results from Paper 3 is reproduced here, rounded to one decimal place to better fit the page width.
Ba = Background tiles, Bl = Blood tiles, Da = Damaged tissue tiles,
Mu = Muscle tissue tiles, St = Stroma tissue tiles, Ur = Urothelium tiles.

Model	Ba	Bl	Da	Mu	St	Ur	Total
TRI-SL-AF	100%	98.6%	89.1%	79.4%	96.4%	98.0%	94.6%
TRI-P-SSL-F	100%	98.6%	90.0%	82.7%	96.1%	98.3%	95.2%
TRI-C-SSL-F	99.9%	96.7%	90.6%	82.5%	95.9%	98.6%	95.1%
TRI-SL-AU	100%	99.9%	87.9%	78.1%	98.1%	99.1%	94.6%
TRI-P-SSL-AU	100%	97.4%	88.2%	82.2%	96.8%	99.5%	94.9%
TRI-C-SSL-AU	100%	98.7%	91.9%	84.7%	95.9%	99.0%	96.0%

5.4.4 Conclusion

The two semi-supervised methods, using both the labeled and unlabeled datasets, outperformed the fully supervised methods, which only use the labeled dataset. The cluster-based self-training method performed best and increased the overall accuracy of the tissue tile classification model from 94.6% to 96% compared to using fully supervised learning with the labeled dataset. In addition, the clustering method generated visually better segmentation images.

5.5 Tissue segmentation comparison

This section summarizes and compares the main results of the three papers related to tissue segmentation. The weighted average F1-score for each class and the total average F1-score for each paper is shown in Figure 5.10.

A straightforward comparison is not possible, as all three papers use different evaluation methods based on different test sets. Paper 1 and 3 evaluated the models on test sets (however, not the same test set), and Paper 2 utilizes a cross-validation approach, where all tiles are used in both training and validation. Hence, the presented comparison in Figure 5.10 must be taken with a grain of salt and not interpreted literally. Still, some insight is given. The right-most columns, indicating the weighted average result, show that the multiscale models in Paper 2 and 3 outperform the single-scale model in Paper 1. For the classes of stroma and muscle, the

two classes with the fewest annotations, it seems like the models can only achieve good results in one of these two classes. Paper 1 and 3 have good results for the stroma class, but have their worst results for the muscle class. Furthermore, Paper 2 has the best results on the muscle class of the three models, but the overall lowest score on the stroma class. All three models have an excellent score on the blood and background classes. And finally, all models have a good score for the urothelium and damaged tissue classes, all well above 90%.

In Paper 3, the TRI architecture from Paper 2 was trained on the $D_{gt}\{train\}$ dataset, and got an accuracy of 94.6% when tested on the $D_{gt}\{test\}$ dataset. However, with the cross-validation scheme in Paper 2, the same architecture got a score of 96.5%. Both papers use the same labels from Dataset B, but in Paper 2, the model is evaluated on all labels and should therefore have a more accurate result of the two. More data is utilized, and it is not surprising that this can improve the results. The claim in Paper 3 that the clustering-based self-training method increases the accuracy of tissue classification should still be valid. Nevertheless, the results of the two papers can not be directly compared.

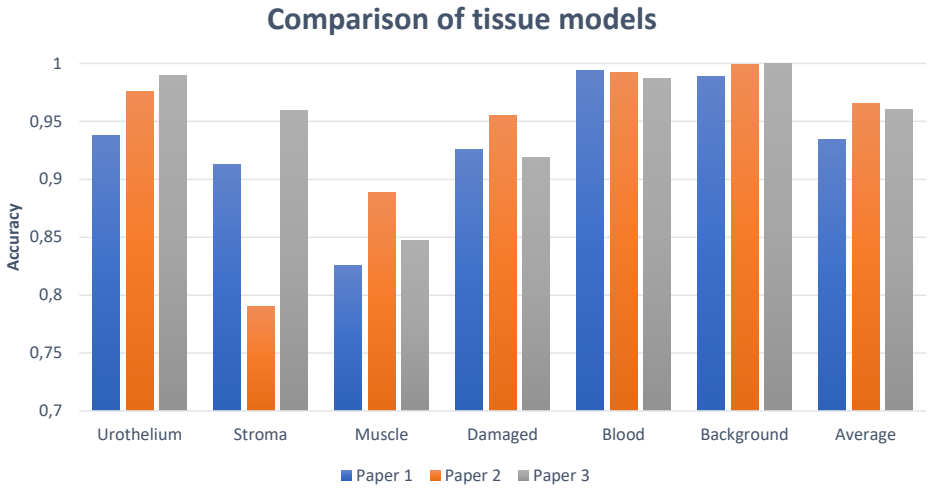


Figure 5.10: Comparison of the best result from each paper on tissue segmentation. Results are for each tissue class and a total average score.

After Paper 3 was published and the work on tissue segmentation was done, one WSI was fully annotated into different tissue types by a pathologist. The annotations have been transformed into a segmentation map

using the same color-mapping function introduced in Paper 2 for easier comparison. The resulting image is considered the ground truth and is shown in Figure 5.11. The same WSI has also been segmented using the best model from Paper 2 and Paper 3, shown in Figures 5.12 and 5.13, respectively.

The predicted images are compared, pixel by pixel, with the ground truth image. The F1-score for each class and a weighted average F1-score for all classes are shown in Table 5.4.

Table 5.4: F1-score for the segmentation maps in Figures 5.12 and 5.13.

Class	Paper 2 (TRI-25x-100x-400x)	Paper 3 (TRI-C-SSL-AU)
Urothelium	0.78	0.79
Damaged	0.50	0.52
Stroma	0.29	0.19
Blood	0.76	0.90
Muscle	0.00	0.00
Background	0.97	0.98
Average	0.92	0.93

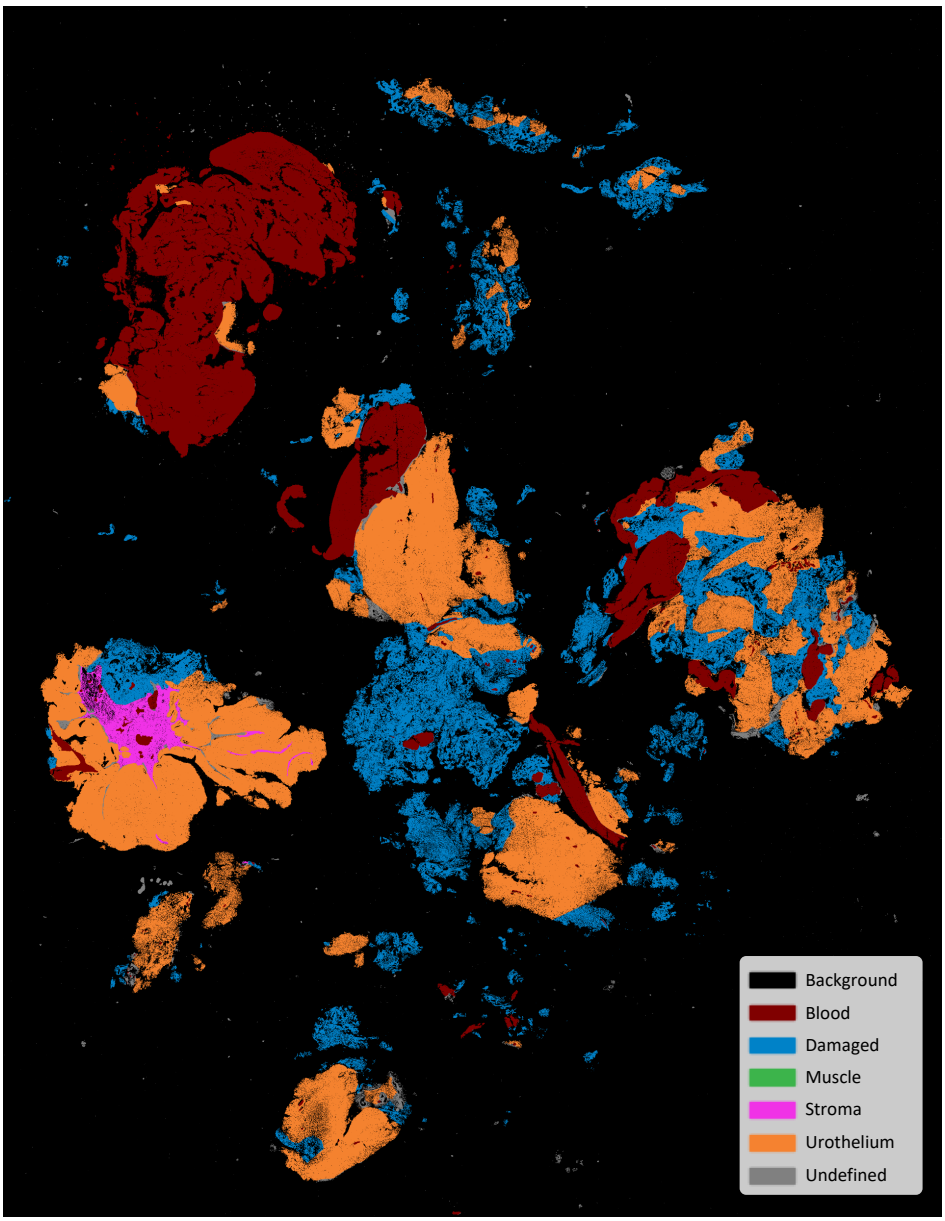


Figure 5.11: Ground truth test image showing regions of all tissue classes in a whole-slide image, annotated by a pathologist.

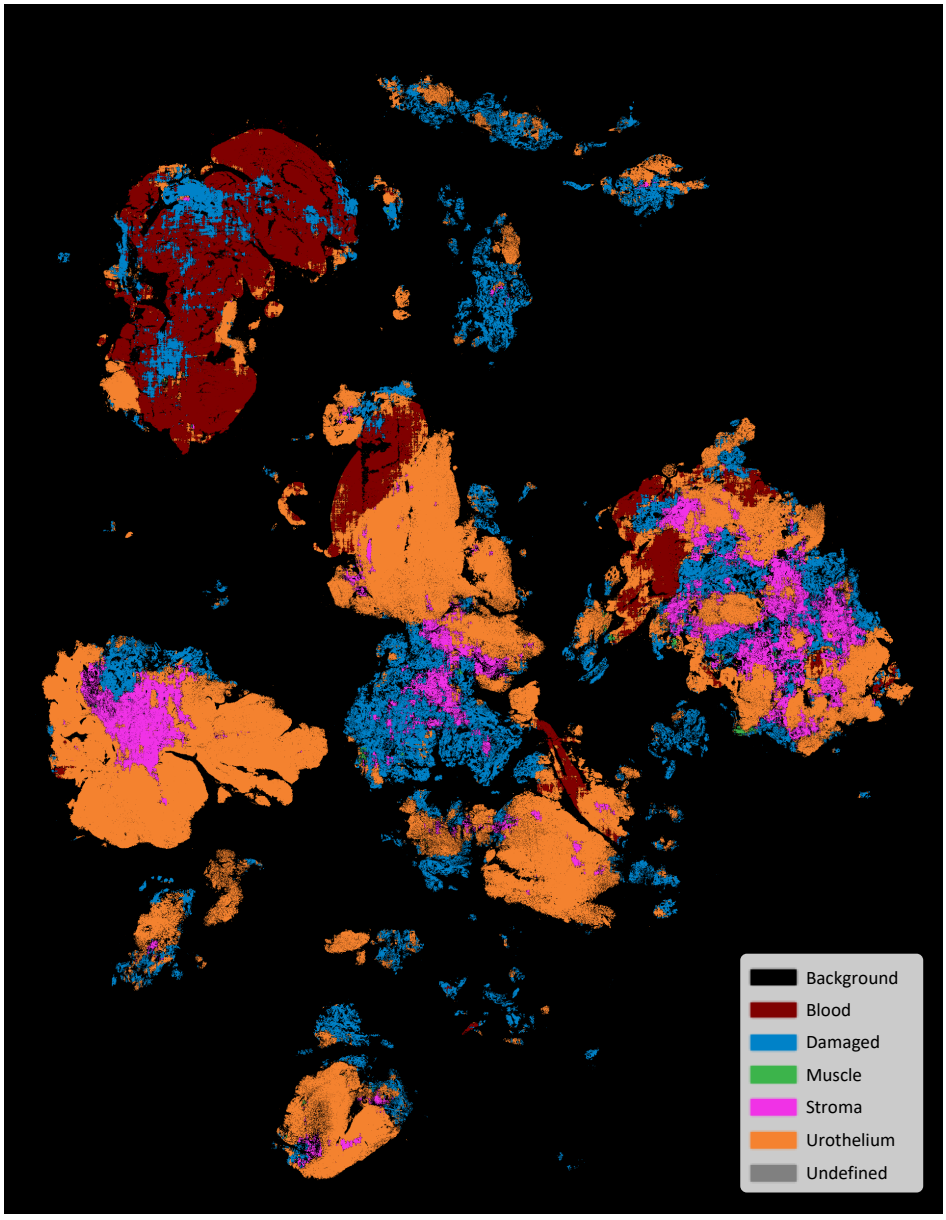


Figure 5.12: The whole-slide image predicted by the best model in Paper 2 (TRI-25x-100x-400x).

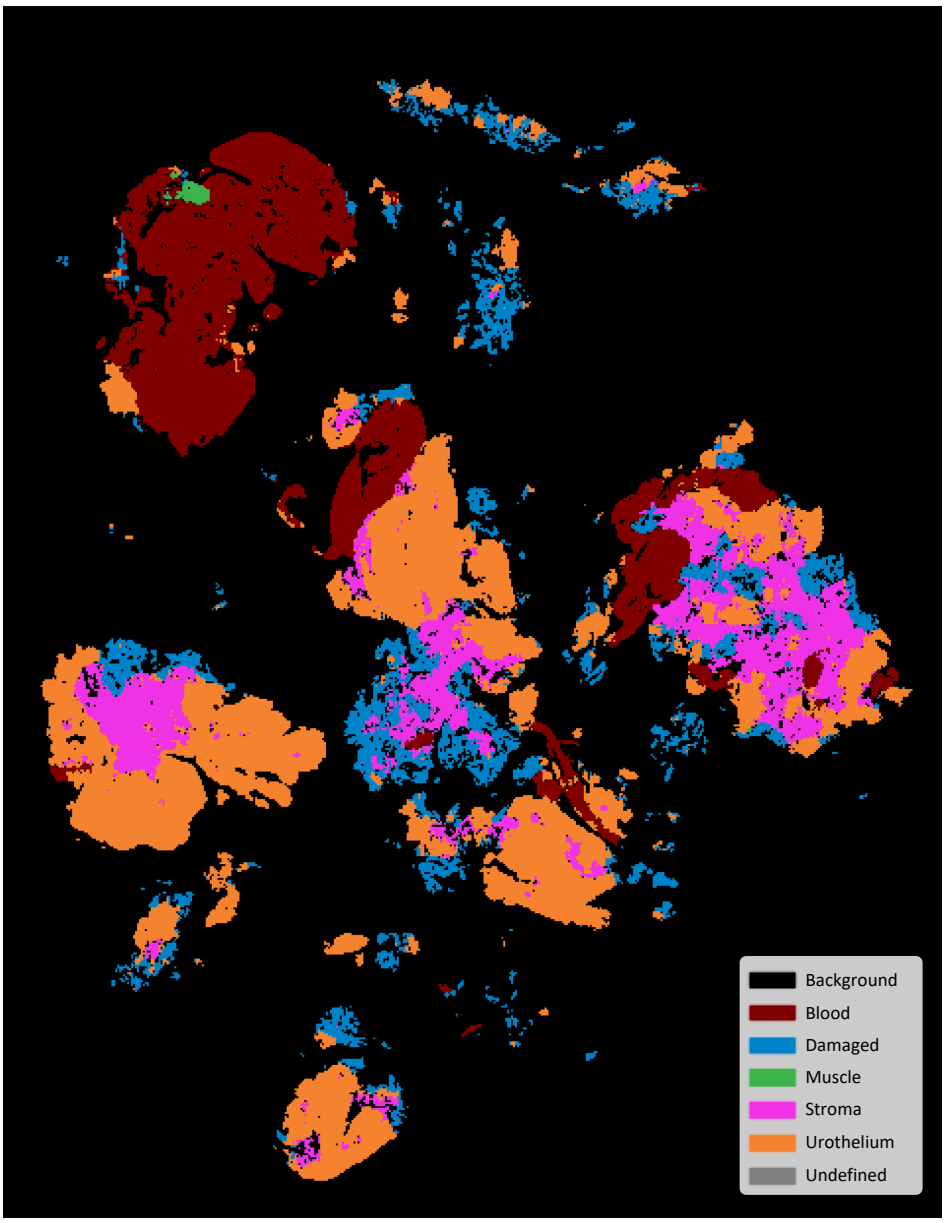


Figure 5.13: The whole-slide image predicted by the best model in Paper 3 (TRI-C-SSL-AU).

Chapter 6

Multilevel tile extraction

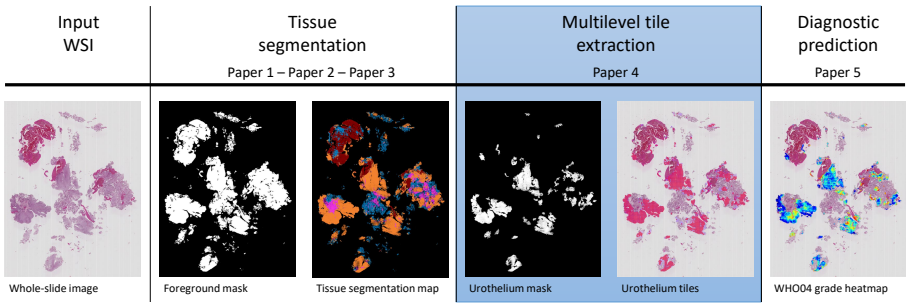


Figure 6.1: An overview of the proposed pipeline, where the topic of this section is highlighted.

The following chapter will present our proposed methods for tile extraction in multilevel gigapixel images; highlighted in blue in Figure 6.1. The main methods and results from Paper 4 are presented. This paper is part of the thesis sub-objective SO_4.

SO_4: Create a reproducible system that automatically extracts tiles from multilevel whole-slide images.

6.1 Paper 4 – Multilevel tile extraction

During the work with tissue segmentation, different ways of dealing with the patches and tiles from the image were tested. However, only straightforward methods were implemented, with few options to control the tile extraction. For instance, in Paper 2, a time-consuming script was made to shift the starting position of the grid in both x- and y-direction to maximize the number of tiles within the ROI. This was implemented because there was

no easy way of controlling the amount of non-ROI in the tiles. During this work, it became clear that even if all papers on WSI processing in the literature use some sort of patching, the process is seldom explained, and it is not very reproducible. Furthermore, to the author's knowledge, no system for visualizing individual tiles from the different magnification levels has been reported.

The main objective of this work is to create a diagnosis system. To be able to create a robust system, most of the WSIs should be utilized in the training/validation/test setup. However, not every single tile in every image should be used. To ensure solid coverage of tiles extracted from each WSI, from all levels in the image pyramid, a sound method for finding and extracting relevant tiles from the WSIs, possibly defined by a ROI mask that can be manually or automatically generated, is needed

In this work, a method for parameterizing and automating the task of extracting tiles from different scales with a ROI defined at one of the scales. The proposed method makes it easy to extract different datasets from the same group of gigapixel images with different choices of parameters, and it is reproducible and easy to describe by reporting the parameters. The method is also used for visualization of tiles from all levels in the pyramid. It is suitable for many image domains and is demonstrated with different parameter settings using histological images from urinary bladder cancer.

6.2 Data material

The methods are demonstrated on histological whole-slide images of urothelial carcinoma, with a corresponding annotation mask generated with the TRI-25x-100x-400x tissue model from Paper 2.

6.3 Method

A 3-level gigapixel image pyramid is depicted in Figure 6.2. Let $\mathcal{S}$ denote the set of levels in a gigapixel image, where $\mathcal{S} = \{0, 1 \dots \ k \dots \ k_{max}\}$. Let $I^k(x, y)$ be an image on level k, where $k \in \mathcal{S}$, and $I^k(x, y)$ has dimension $N_k \times M_k$. A binary mask $B^k(x, y)$ representing the ROI at level k, also have the same dimension $N_k \times M_k$.

A tight grid of non-overlapping tiles is superimposed on the image, where the upper-left corner of the grid starts at the image coordinate $(0, 0)$. Let

the upper-left coordinates of a tile in image $I^k(x, y)$ be denoted (x^k, y^k), and let the tile have dimension $L_T^k \times L_T^k$. A *valid tile* is a tile constrained by the annotation mask on level $I^\alpha(x, y)$, where $\alpha \in \mathcal{S}$ and refers to the constraining level.

Once a valid tile is found on level α, tiles from $I^\beta(x, y) \forall \beta \in \mathcal{S}_\beta$ are found and referred to as *corresponding tiles*, where $\mathcal{S}_\beta$ is a subset of $\mathcal{S}$ and contains the remaining levels to extract tiles. I.e., to extract tiles from all three levels in Figure 6.2 with level 0 as the constraining level, we would specify $\alpha = 0$ and $\mathcal{S}_\beta = \{1, 2\}$.

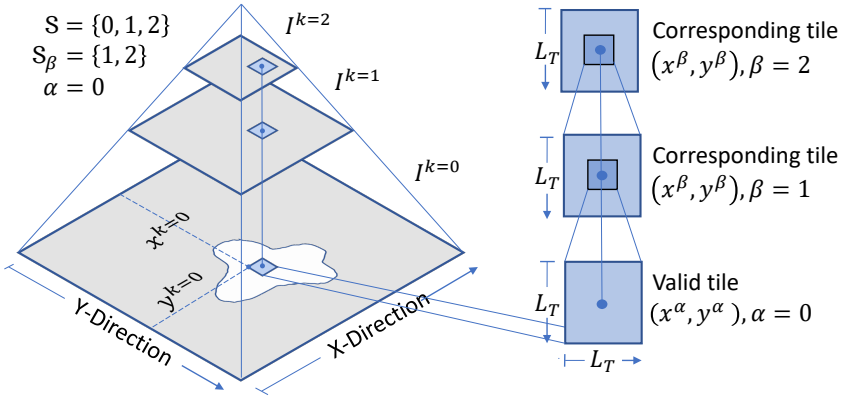


Figure 6.2: The left-hand side displays a 3-level image pyramid, and the right-hand side shows a valid tile found on level $\alpha = 0$ and the two corresponding tiles for level $\beta = 1$ and 2. L_T is the same in all levels.

τ specifies the current level a tile is on, and is necessary to maintain the generalization of the equations. τ may, or may not, be equal to α, as it is possible to have, e.g., a tile at $\tau = 2$, and check if it is valid on level $\alpha = 1$

The parameter ϕ is defined to parameterize the overlap-ratio between the tile and the ROI at a chosen magnification level. $\phi \in [0, 1]$, where $\phi = 1$ means the ratio between the tile area and ROI must be 1 for the tile to be valid, i.e., the entire tile must be inside the ROI. $\phi = 0$ would correspond to the ROI being ignored, and all tiles within the image would become valid, whereas $\phi = 0.8$ means that 80% of the tile at that level is required to be inside the ROI for it to become valid.

It is necessary to define the size of level k relative to the base level as a ratio $R_k = \frac{N_k}{N_0}$. The size-relationship between an arbitraty level k_1 and level k_2 is found as $R_{k_1 k_2} = \frac{R_{k_1}}{R_{k_2}}$

The parameterization of the tile-extraction is defined by α, ϕ L_T^k, $\mathcal{S}_\beta$, and the starting coordinate of the tile grid. The method is divided into three parts, presented here.

Part 1 - Parameterizing of overlap-ratio with ROI

The input to the method is a histological WSI. These images contain a lot of unwanted tissue types, like blood and damaged tissue. To avoid such regions, we use an annotation mask as a reference, for example, the urothelium mask in Figure 6.1. It is a binary mask and contains 1's for the urothelium tissue and 0's everywhere else.

We want to allow some tile-area outside the ROI, as this can be useful if the region contains voids of background, like the example tile in Figure 6.3-A. Also, in some cases, the tissue border may reveal diagnostic information, and the only way to include the border in a diagnostic dataset is to allow for some background when extracting tiles. An example of such a tile is shown in Figure 6.3-B.

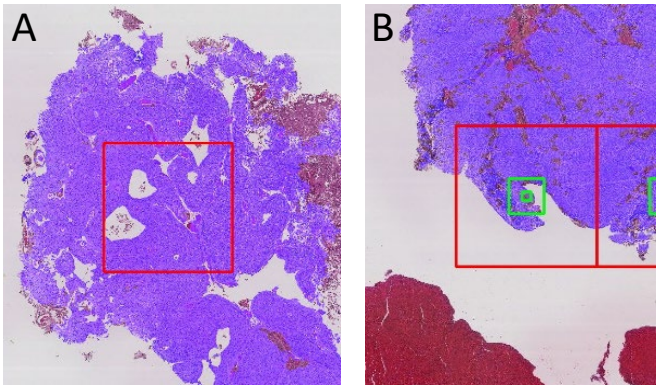


Figure 6.3: Examples of tiles positioned to include both ROI and non-ROI areas. In image A, the ROI contains voids, and in image B, the tile covers the edge of the tissue.

The parameter ϕ controls how much non-ROI area to allow in the extracted tiles. To determine if the tile (x^α, y^α) is a valid tile, it must satisfy the following condition:

$$\left[\sum_{i=x^\alpha}^{x^\alpha+L_T^\alpha\cdot R_{\alpha\tau}}\sum_{j=y^\alpha}^{y^\alpha+L_T^\alpha\cdot R_{\alpha\tau}} B^\alpha(i,j)\right] \geq (L_T^\alpha \cdot R_{\alpha\tau})^2 \cdot \phi \tag{6.1}$$

In Figure 6.4, a tile is positioned at location (x^α, y^α), containing multiple voids of background. For $\phi = 1$ this tile would not satisfy Equation 6.1 and thus be discarded. However, for $\phi <\approx 0.7$ the tile would satisfy the condition and thus be considered a valid tile and be included.

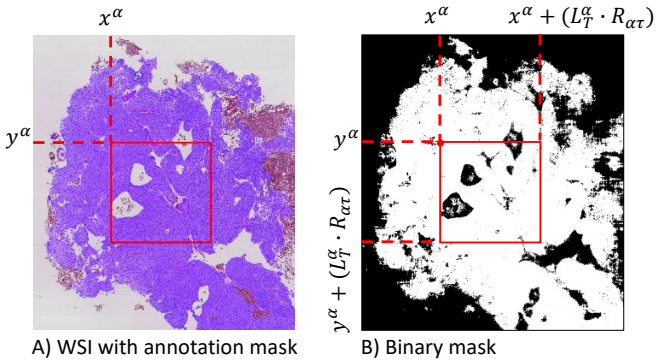


Figure 6.4: A) A WSI image with an annotation mask (semi-purple) and a red tile. B) Binary mask with the same tile, showing the minimum and maximum coordinates for the tile.

Part 2 - Finding corresponding tiles on levels $\beta \in \mathcal{S}_\beta$

In part 2, we want to use the valid tiles from part 1, and find the corresponding tiles on level beta.

Tiles are usually referenced by the upper-left corner. However, when going between levels, it is convenient to use the center point of the tile, as visualized in Figure 6.2.

We do this by adding half of the tile's length L_T and then transform this position to the corresponding level by multiplying with the ratio between the two levels. After transformation, we need to return to the upper-left corner by subtracting half of the tile's length.

This can be generalized into an expression for the upper-left coordinate of corresponding tiles, one for X- and one for Y-coordinate (only X-coordinate is shown here).

$$x^\beta = (x^\tau + \frac{L_T^\tau \cdot R_{\alpha\tau}}{2}) \cdot R_{\tau\beta} - \frac{L_T^\tau}{2} \forall \beta \in \mathcal{S}_\beta \tag{6.2}$$

Part 3 - Visualize all tiles on one sub-image

Typically extracted tiles are spread onto k_{max} different images, and due to the enormous size of the highest resolution image in the pyramid, we cannot use these coordinates directly for visualization. Thus, we want to transform all tiles onto the smallest image in the pyramid, $I^{k_{max}}(x, y)$, which is suitable for viewing on a monitor.

For this, we need to alter Equation 6.2 slightly. First, $R_{\tau\beta}$ is changed to $R_{\tau k_{max}}$ to transform all coordinates to level k_{max}. Second, the last term L_T is changed to $L_T \cdot R_{\beta k_{max}}$ because we need to rescale the tile's apperent length so that it covers the same physical area on both level β and level k_{max} for visualization.

To formalize these operations for all levels, the following equations will transform the upper-left coordinate of a tile on level τ into the upper-left coordinate of a tile on level k_{max}.

$$x_\beta^{k_{max}} = (x^\tau + \frac{L_T^\tau \cdot R_{\alpha\tau}}{2}) \cdot R_{\tau k_{max}} - \frac{L_T^\tau \cdot R_{\beta k_{max}}}{2} \forall \beta \in \mathcal{S}_\beta \tag{6.3}$$

The transformation, or visualization, of tiles from all levels onto level k_{max} is illustrated in Figure 6.5, where the left-hand side of the figure represents the input coordinates to Equations 6.3, and the right-hand side represents the output.

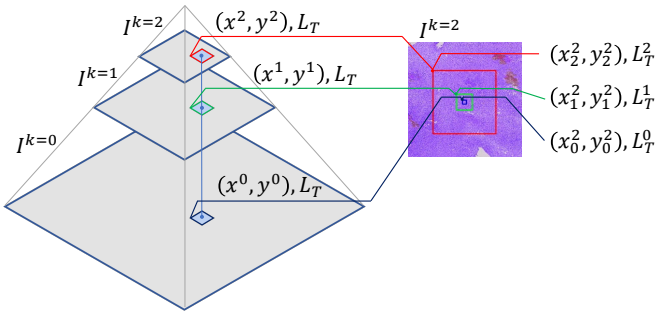


Figure 6.5: Tiles from all levels are combined on the highest level to be visualized next to each other.

The proposed methods are primarily used for preprocessing to extracting tiles used to train a deep learning model, as shown in Figure 6.6-A. However, part 3 is for visualization of the extracted tiles, and this method can also be

used in post-processing. E.g., the method can be used after a deep learning model has classified the tiles, either by plotting the tiles using colors to indicate correct and incorrect predictions, as depicted in Figure 6.6-B, or by using the model's probabilistic output score to generate a heatmap, as shown in Figure 6.6-C.

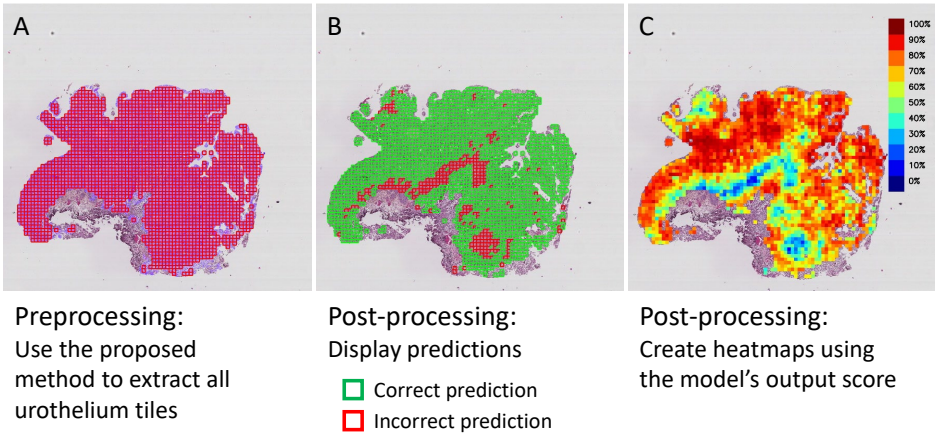


Figure 6.6: The proposed methods in this paper can be used for both preprocessing and post-processing. A) shows an example of how the method finds and extract tiles from WSIs, B) tiles are colored according to the model's prediction, highlighting correct and incorrect model predictions, and lastly, C) the model's probabilistic score is used to color the position of each tile and create a heatmap.

6.4 Result

To demonstrate the method, an example WSI is used together with a urothelium mask and a stroma tissue mask. The stroma mask is not used in the automatic grading system in this thesis; but is included in this demonstration to show the method's ability to include several tissue classes. In Figures 6.7, 6.8 and 6.9, the proposed methods are used to find tiles of size 256×256 pixels in the WSI, with three different values of alpha. ϕ is set to a relatively low value of 0.4 for all three figures, but is easiest to see in Figure 6.7 by the amount of non-ROI regions in the red, valid tiles.

The red tiles indicate the constrained *valid* tile, i.e., at $k = \alpha$, and the green tiles are the *corresponding* tiles found from the other levels. Note that the red constraining tiles never overlap, but the green tiles may do so.

Also, the red tiles are restricted to lay within the ROI, but the green tiles may cover an area beyond the region border.

In Figure 6.9, tiles are found on the $I^{k=0}$ image, the largest image in the pyramid. This results in the most tiles found, 9 362 tiles in this instance. Whereas in Figure 6.7, we are finding tiles on the smaller $I^{k=2}$ image in the pyramid. Because the dimension of this image is smaller than that of the $I^{k=0}$ and the tile size L_T is the same, the apparent size of the tiles is larger. Consequently, there is room for fewer tiles, and only 29 valid tiles are found.

The advantage of setting $\alpha = 0$ is a large dataset; however, this dataset will contain redundant pixel information on levels 1 and 2. A smaller dataset is found by setting $\alpha = 2$, but each tile will be unique on all levels in the image pyramid.

For a given image $I^k(x, y)$ with a binary annotation mask $B^k(x, y)$, by specifying α, ϕ, L_T^k, $\mathcal{S}_\beta$, and the grid starting coordinates, all extracted tiles from all levels in the image is uniquely determined, and the process is repeatable and reproducible.

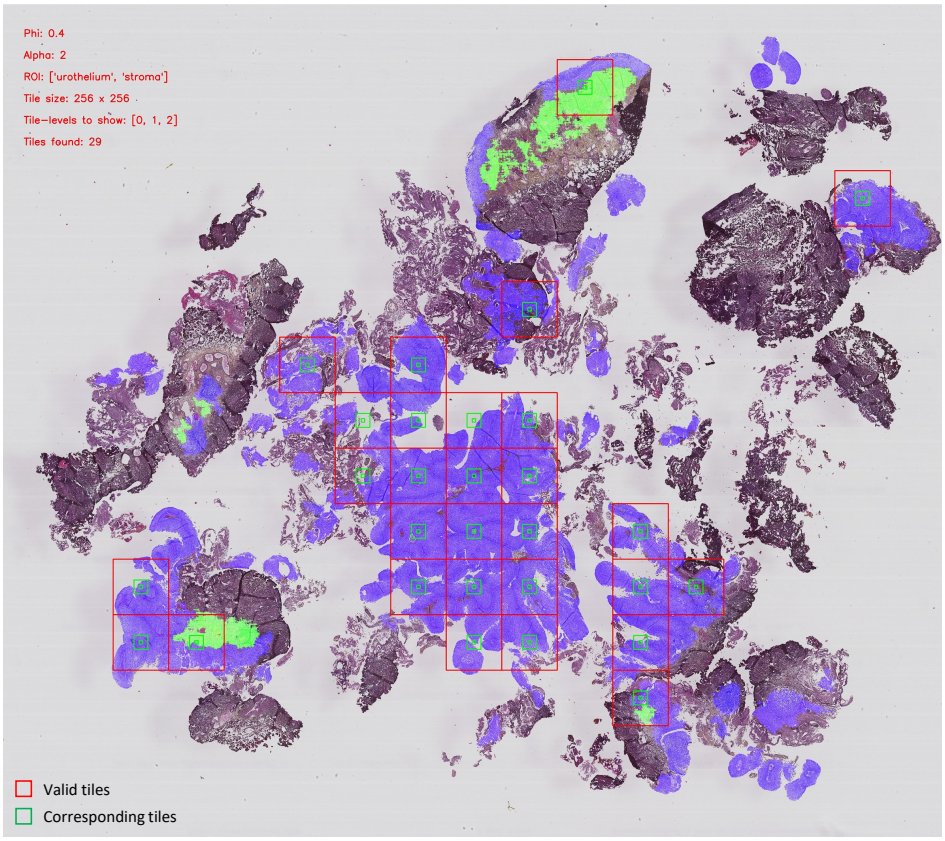


Figure 6.7: Visualization of all valid and corresponding tiles found in the WSI. The urothelium mask is shown as purple, and the stroma mask in green. In this example, $\alpha = 2$, $L_T = 256$, and $\phi = 0.4$. Tiles from all levels are shown. These values are just for demonstration and are not used to extract the diagnosis dataset.

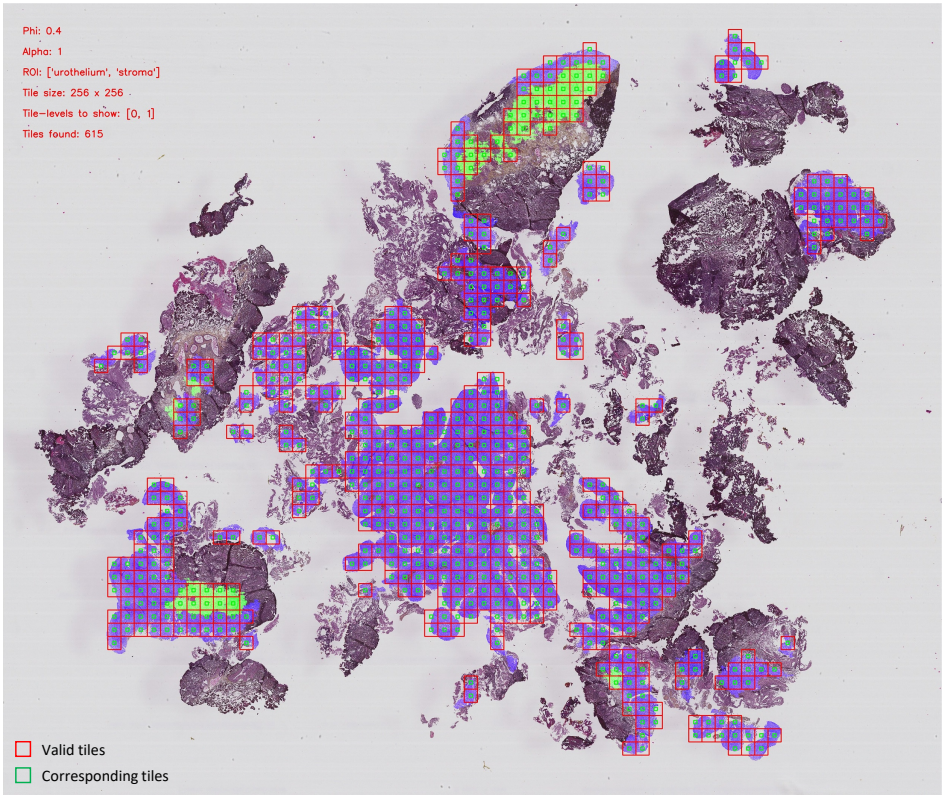


Figure 6.8: Visualization of valid and corresponding tiles found in the WSI. The urothelium mask is shown as purple, and the stroma mask in green. In this example, $\alpha = 1$, $L_T = 256$, and $\phi = 0.4$. Tiles from levels 0 and 1 are shown, but tiles from all levels are extracted and can be used in a dataset. These values are just for demonstration and are not used to extract the diagnosis dataset.

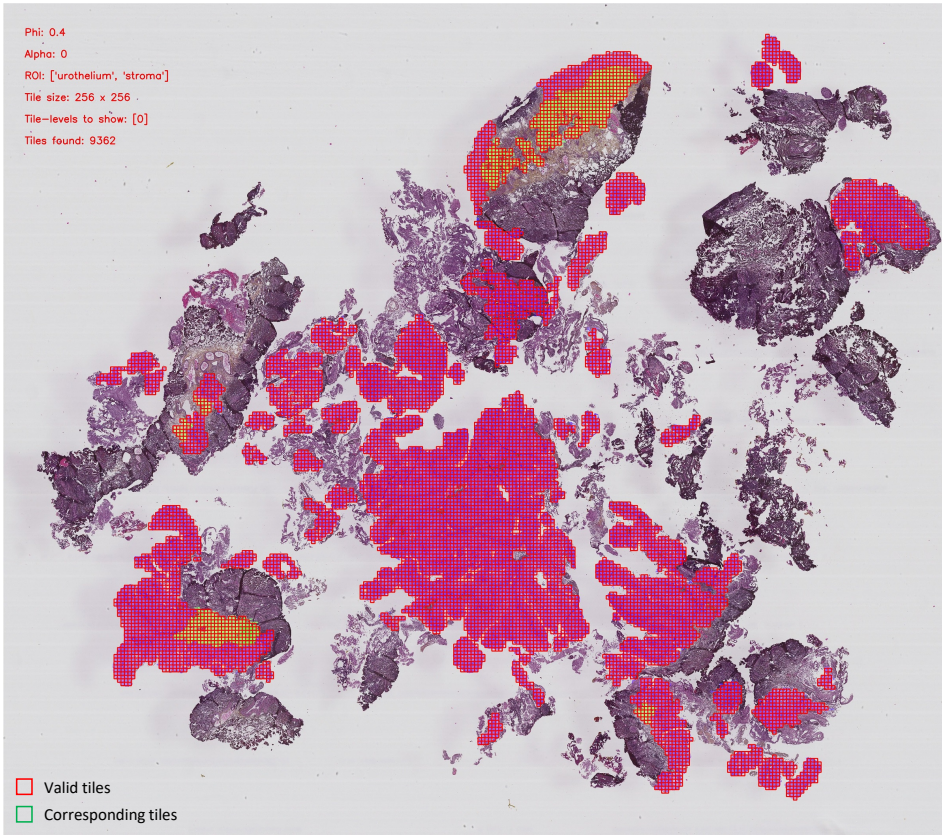


Figure 6.9: Visualization of valid tiles found in the WSI. The urothelium mask is shown as purple, and the stroma mask in green. In this example, $\alpha = 0$, $L_T = 256$, and $\phi = 0.4$. Only tiles from level 0 are shown, but tiles from all levels are extracted and can be used in a dataset. These values are just for demonstration and are not used to extract the diagnosis dataset.

6.5 Conclusion

The proposed method for parameterized tile extraction was very useful in the work in this thesis. Using this method to patch up and extract datasets from ROI makes it reproducible and easy to report, which can be useful in other types of gigapixel images as well as WSI applications. The additional ability to use the method to generate heatmaps also proved to be a valuable and important factor in the work of the diagnosis system.

Chapter 7

Diagnostic prediction

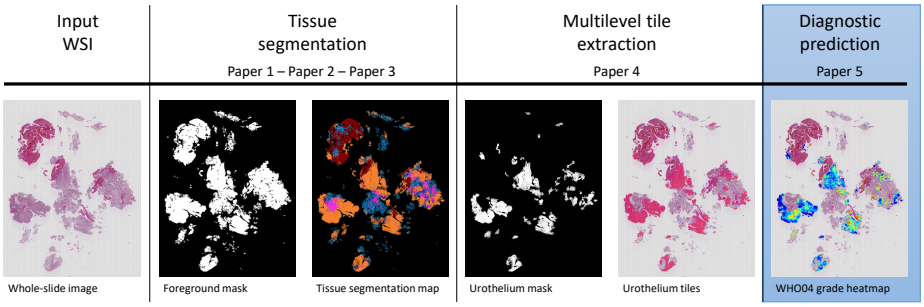


Figure 7.1: An overview of the proposed pipeline, where the topic of this section is highlighted.

The following chapter will present the proposed method for automatic grading of urothelium carcinoma whole-slide images, highlighted in blue in Figure 7.1. First, an overview of the contribution is given, then the main methods and results from Paper 5 are presented. This paper is part of the thesis's main objective O_1.

O_1: Create a system for automated grading of urothelial carcinoma slides.

7.1 Contribution overview

Through the Stavanger University Hospital, we have access to a large dataset of bladder cancer WSIs. In addition, each WSI has been graded and staged by a pathologist. We want to use the tissue segmentation model from Paper 2 to extract urothelium tissue from all WSIs in the dataset and use the slide-level WHO04 grades as labels to train an automatic diagnosis system. The output of such a system could be heatmaps, visualizing locations of

low- and high-grade carcinoma in the WSIs, in addition to slide-level grade predictions.

Successful implementation of a grading system has some potential applications. For example, where the tissue segmentation images can guide the pathologists to the diagnostic relevant urothelium areas in the WSIs, the addition of WHO04 grading heatmaps could help narrow this further down to the most *severe* urothelium regions, making their workflow even more efficient.

In a clinical setting, the automated diagnosis system could run without supervision from the clinician and predict the slide-level grade of newly scanned glass slides. The predicted grade could then be used to prioritize high-grade patients for earlier examination. Also, it can be used as input to an automatic prognostic tool and output a measure of the patient's overall clinical outcome, such as the risk of recurrence, 1-yr and 5-yr survival rate, and mortality. In the future, it is also a possibility to use it in an automatic system that predicts how a patient will respond to a given treatment and therapy program.

The current study's main contribution is the demonstration of how two systems (tissue and diagnosis) are combined into one fully automatic pipeline for tissue segmentation, generating WHO04 heatmaps and providing a slide-level grading.

7.2 Paper 5 – Diagnostic prediction

In this paper, we propose a pipeline called $\text{TRI}_{\text{grade}}$ for automatically grading WSI according to the WHO04 grading system. The pipeline is depicted in Figure 7.2. The system will identify diagnostic relevant regions in the WSI and collectively predict the grade. The proposed system uses the $\text{TRI}_{\text{tissue}}$-model as a first-stage network for preprocessing the WSI to find regions of urothelium tissue. Next, the extracted urothelium tissue is then fed through a second-stage network, called $\text{TRI}_{\text{WHO04}}$-model, for automatic WHO04 grading. The $\text{TRI}_{\text{grade}}$ pipeline will output a tissue segmentation map, a WHO04 grading heatmap, and a slide-level WHO04 grade prediction. A depiction of the pipeline is shown in Figure 7.2 and explained in detail below.

In Paper 2, three architectures were tested (MONO, DI, and TRI) with both frozen and unfrozen weights. The best result was achieved with the

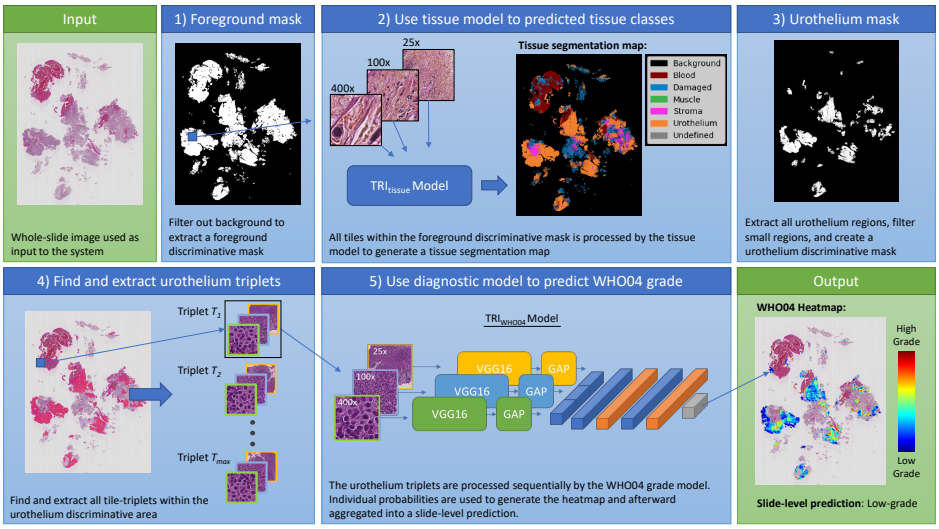


Figure 7.2: This figure presents the pipeline for our proposed system, $\text{TRI}_{\text{grade}}$. **Input)** A WSI of urothelial carcinoma is used as input. **1)** A foreground discriminative mask is found by evaluating the pixel intensity values and used as a reference to extract tiles from the WSI. **2)** The $\text{TRI}_{\text{tissue}}$-model is used to generate a tissue segmentation map. **3)** The urothelium regions are used to create a urothelium discriminative mask. **4)** Using the urothelium mask, triplets consisting of tiles from three magnification levels are extracted from the input WSI. **5)** The urothelium triplets are fed sequentially to the $\text{TRI}_{\text{WHO04}}$-model, which outputs a probabilistic score for the two classes, low- and high-grade carcinoma. **Output)** The system will output a WHO04 grade heatmap and a slide-level WHO04 prediction.

frozen TRI-model, utilizing all scales, and achieved an average F1-score of 97.6% for the urothelium class.

Based on this result, we continued with the TRI-model and VGG16 as feature extractors in Paper 5. We have not evaluated the MONO- or DI-models on the diagnostic data. The model referred to as TRI-25x-100x-400x in Paper 2 is in the current paper referred to as the $\text{TRI}_{\text{tissue}}$-model. It is used for tissue extraction as shown in Figure 7.2. The name, architecture, and base model have also been carried over to this paper and are the basis for the $\text{TRI}_{\text{WHO04}}$-model we propose here.

Because only the TRI-model is used throughout this paper, tiles will always be extracted from all levels. For improved readability, we define these tiles as a *triplet*. A triplet is denoted T_i and is defined as a set of three tiles extracted from a WSI at three different magnification levels (25x, 100x, and 400x). Let $\mathcal{T}$ denote a set of triplets in a WSI, where

$\mathcal{T} = \{T_1, T_2 \dots\ T_i \dots\ T_{max}\}$, and the number of elements in the set is given by the cardinality $|\mathcal{T}|$. Example of tissue triples are shown in Figure 2.2, examples of low- and high-grade triplets are shown in Figure 2.3, and finally, examples of how triplets are extracted from the image pyramid is shown in Figures 4.3 and 6.2.

The CNN-based model assigns a prediction score to every tile. These predictions are used to create a heatmap showing which regions were predicted with low- or high-grade carcinoma. The final decision can further be aggregated from the predictions into a slide-level prediction.

The proposed method is inspired by the behavior of a pathologist by combining global context information and local details by utilizing a multiscale model architecture.

7.3 Data material

Three datasets were used together with the slide-level labels as ground truth, as shown in Figure 4.7. Dataset E was used to train the model, consisting of 220 WSIs, and the 30 WSIs in Dataset F were used as validation dataset. The method was evaluated on 50 WSIs from Dataset G. In addition, a pathologist has carefully annotated low- and high-grade regions in 14 of the 50 WSIs in Dataset G. The 14 WSIs are a sub-set of the test set and are referred to as the *segmentation test set* and will be used to evaluate the low- and high-grade segmentation performance of the best $\text{TRI}_{\text{WHO04}}$-model.

All WSIs were preprocessed with the $\text{TRI}_{\text{tissue}}$-model to create a tissue segmentation map. From this segmentation, the urothelium regions were extracted, filtered to suppress noise, and used to generate a binary urothelium discriminative mask.

Using the urothelium mask as a reference, the method for tile extraction described in Paper 4 was used to extract all the urothelium tiles from the WSIs.

From the 220 WSIs used for training, five datasets were extracted with a different number of triplets extracted from each WSI. A set of N triplets was selected randomly from the predicted urothelium regions in each WSI, where N was set to 250, 500, 1 000, 3 000, and 5 000 in different experiments.

An augmentation strategy was employed for WSIs with fewer than N triplets. In these WSIs, randomly selected triplets were rotated, and vertical/horizontal mirrored until the desired number of N triplets was

reached, or the maximum of eight times augmentation was reached. The aim of this process is for each WSI to contribute equally, or as close as possible, to the number of triplets specified by N. No augmentation was performed on the validation or test datasets.

Each extracted tile inherits the slide-level WHO04 grade as its label. This is not ideal, as high-grade slides may contain regions with low-grade tissue. Consequently, all the extracted datasets are thus regarded as weakly labeled due to the inaccurate labels, which is consistent with what is called a weak label in many tasks [27]. On the other hand, the segmentation test set is considered strongly labeled.

7.4 Method

7.2 contains five steps explained here. The input to the pipeline consists of a WSI. First, in step 1, a foreground discriminative mask is found on the 400x level by evaluating the pixel intensity values as grey background or not. Using the foreground mask as a reference, tiles of size 128 $\times$ 128 pixels were extracted from the WSI, such as the inner 16 $\times$ 16 pixels were being classified for each tile. Three tiles were extracted in the WSI (25x, 100x, and 400x) for each location, forming a triplet.

In step 2, triplets are sequentially fed into the $\text{TRI}_{\text{tissue}}$-model from Paper 2. This model will evaluate the triplets and predict the tissue class for the current triplet. After classifying all triplets, a segmented tissue map is created, visualizing all tissue regions in the WSI. This tissue map can also be presented to the clinician to help guide them more efficiently to the specific tissue types in the WSI.

From the generated tissue map, all urothelium regions are extracted in step 3. Small regions are filtered to suppress noise, and a urothelium discriminative mask is created on the 400x level. In step 4, a grid of non-overlapping tiles is overlayed on the WSI at the 400x level, this time using tiles of dimension 256 $\times$ 256 pixels. The individual tiles in the grid are checked against the discrimination mask. If 80% or more of a tile lay within the discriminate mask, the position is saved, while the remaining tiles are discarded. For the validation and test sets, triplets from all the saved positions are extracted. Whereas for the training set, N randomly selected triplets are extracted from the saved positions. If fewer than N positions are saved, the augmentation strategy explained in the data material section

is employed. The total number of extracted triplets for each dataset is shown in Figure 4.7, as well as in Paper 5.

Tiles are extracted from the WSI as described in Paper 4. The parameters used are the tile size parameter $L_T = 256$. The overlap-ratio between a tile and the discriminative mask is set to 80%, which corresponds to a value of $\phi = 0.8$. Tiles are checked at the 400x level by setting $\alpha = 0$, and the corresponding tiles in the triplets are found at level 25x and 100x, i.e., $\mathcal{S}_\beta = \{1, 2\}$. The binary mask B^k is set as the urothelium discriminative mask, and the starting coordinate of the grid is at position $(0, 0)$. With these parameters and the methods described in Paper 4, extraction of the triplets in the WSIs is repeatable and reproducible.

In step 5, the extracted urothelium triplets are fed to the $\text{TRI}_{\text{WHO04}}$-model, which outputs a probabilistic score for the two classes, low- and high-grade carcinoma. Finally, all scores are used to generate a heatmap which is overlayed on the WSI, and the aggregated predictions are measured against the decision threshold D_t to get the final slide-level prediction.

For the heatmaps, only the model's probabilistic score for the high-grade class is used to generate the heatmaps. However, because there are only two classes, a low probabilistic score of the high-grade class implicitly means a high score for the low-grade class. I.e., red highlighted regions in the heatmaps are associated with the high-grade class, and blue highlights indicate the low-grade class.

Model architectures

A block diagram of the $\text{TRI}_{\text{WHO04}}$-model architecture is depicted in Figure 7.3, and a block diagram of the $\text{TRI}_{\text{tissue}}$-model is depicted in Figure 5.6. The architecture is almost the same, but the $\text{TRI}_{\text{tissue}}$-model has six output classes instead of two.

Tile-level prediction

When a triplet T_i is fed to the $\text{TRI}_{\text{WHO04}}$-model, the model outputs a list of probabilities for the two classes, low- and high-grade. These probabilities are denoted as $[p_l^i, p_h^i]$, and ordered such as the low-grade class is located at position 0, and the high-grade class at position 1. To find the class with the largest predicted probability, the argmax function is used.

$$c_i = argmax([p_l^i, p_h^i]) \tag{7.1}$$

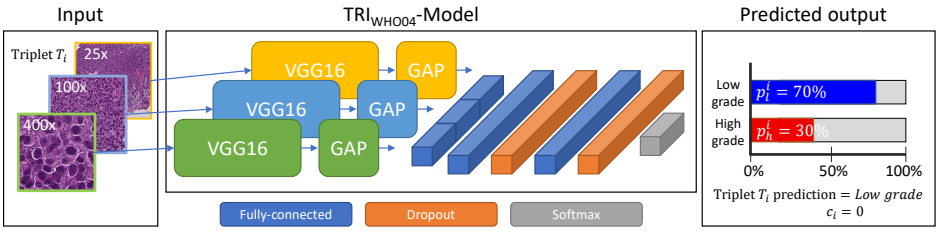


Figure 7.3: Architecture of the TRI$_{WHO04}$-model. Three separate VGG16 networks are used to extract features from each magnification scale. The global average pooling layer (GAP) is used to flatten the features into feature vectors, which are concatenated. The classification network consists of fully-connected layers and dropout layers. The output uses a softmax activation function to predict the input tiles to the two classes, low-grade and high-grade carcinoma.

Where c is given the positional value of the class with the greatest probability score of a triplet at position i. For the triplet in Figure 7.3, the output prediction is $[0.7, 0.3]$, which gives a value of $c = 0$.

Slide-level prediction

In addition to predicting the individual triplets, we also output a WHO04 slide-level prediction. A pathologist will often assign the worst case to a slide during a clinical examination, meaning that if a high-grade region exists in the WSI, the WHO04 grading should be high-grade. However, we must assume some misclassification in the WSI from both the TRI$_{tissue}$-model and TRI$_{WHO04}$-model, so there must be a minimum amount of high-grade triplets before the slide-level prediction becomes high-grade, and we would like to find a decision threshold, D_t, which maximizes correct prediction of the WSIs.

By summing over c_i, the number of triplets predicted as high-grade is counted since triplets predicted as low-grade is at index 0 and therefore not adding to the sum. Thus, by dividing by the total number of triplets in the WSI, we get the ratio of high-grade triplets referred to as R_{high} in this paper:

$$R_{high} = \frac{1}{|\mathcal{T}|} \sum_{i=1}^{|\mathcal{T}|} c_i \tag{7.2}$$

If R_{high} exceeds the decision threshold D_t, the slide is given the slide-level prediction of high-grade; else, it is considered low-grade.

$$\hat{Y} = \begin{cases} \text{High-grade}, & \text{if } R_{high} \geq D_t \\ \text{Low-grade}, & \text{otherwise} \end{cases} \tag{7.3}$$

An algorithm is used to determine the optimal threshold value D_t by looping through and testing all threshold values between 0-50%. The decision threshold D_t is chosen as the threshold with the highest score, or, if more than one value yielded the same maximum result, the average integer value is selected as the decision threshold D_t. The algorithm and a description of it, is shown in the paper.

Experiments

Two experiments were conducted in the paper, where the first experiment is for evaluating the slide-level predictions and the second experiment for tile-level predictions. First, ten identical versions of the $\text{TRI}_{\text{WHO04}}$-model was trained on ten different training datasets. All ten training dataset used the same 220 WSIs, but N triplets were extracted per WSI, where N was set to 250, 500, 1 000, 3 000, and 5 000. Each of these datasets was trained with and without augmentation. This experiment was conducted to see if it is preferable to utilize more of the available urothelium data from each WSI as training data at the cost of additional training time or if a smaller dataset could perform equally well.

The predictions from the individual triplets were aggregated into a slide-level prediction of the WHO04 grading. A decision threshold D_t was found for each model using Algorithm 1; then, equation 7.3 was used to provide the final predicted grade.

Each model was trained until an early stopping criterion monitoring the validation dataset loss stopped the training. Each model was then evaluated on the test set. Next, the best model from the first experiment was used to evaluate the segmentation test set, and the model is used to create heatmaps, visualizing low- and high-grade regions in the WSI.

7.5 Result

Ten models were trained for the slide-level grading experiment, where the best model correctly predicted 45 of the 50 WSIs in the test set, achieving an accuracy of 90%. Slide-level results for all models are listed in Table 7.1.

Table 7.1: Slide-level prediction results for automatic WHO04 grading tested on the 50 WSIs of the test set. Precision, recall, and F1-score is the weighted average score for the two classes across all 50 WSIs in the test set. D_t is the decision threshold found using Algorithm 1. Columns for trained epochs, time per epoch, and training time has been omitted from the table to better fit the pages width. Full table available in Paper 5.

Model	Precision	Recall	F1-Score	D_t
$\text{TRI}_{\text{WHO04}}$-250	0.86	0.84	0.84	49
$\text{TRI}_{\text{WHO04}}$-250-AUG	0.89	0.86	0.85	47
$\text{TRI}_{\text{WHO04}}$-500	0.89	0.86	0.85	43
$\text{TRI}_{\text{WHO04}}$-500-AUG	0.77	0.76	0.76	49
$\text{TRI}_{\text{WHO04}}$-1000	0.83	0.82	0.82	49
$\text{TRI}_{\text{WHO04}}$-1000-AUG	0.80	0.80	0.80	49
$\text{TRI}_{\text{WHO04}}$-3000	0.89	0.86	0.85	49
$\text{TRI}_{\text{WHO04}}$-3000-AUG	0.78	0.78	0.78	49
$\text{TRI}_{\text{WHO04}}$-5000	0.85	0.84	0.84	48
$\text{TRI}_{\text{WHO04}}$-5000-AUG	**0.92**	**0.90**	**0.90**	48

The best model, $\text{TRI}_{\text{WHO04}}$-5000-AUG, was further evaluated on the smaller segmentation test set, where it achieved an average F1-score of 0.91 for both the low-grade and high-grade classes.

A direct comparison of results with others reported in the literature is not straightforward, as the experiments performed in this paper are conducted on a private dataset, which is often the case in many medical applications. In the paper, the results of the $\text{TRI}_{\text{grade}}$ pipeline were compared to the work of Jansen et al. [59]. The works are performed on different datasets but are otherwise quite similar; they are based on an NMIBC dataset of similar size (328 WSIs from 232 patients vs. our dataset of 300 WSIs from 300 patients), a similar split of the dataset into training, validation, and test, and the use of the same labels (WHO04). The results are shown in table 6 in Paper 5. We achieve better results on all metrics, and with 45 of the 50 WSIs correctly predicted, we achieve an accuracy of 90%.

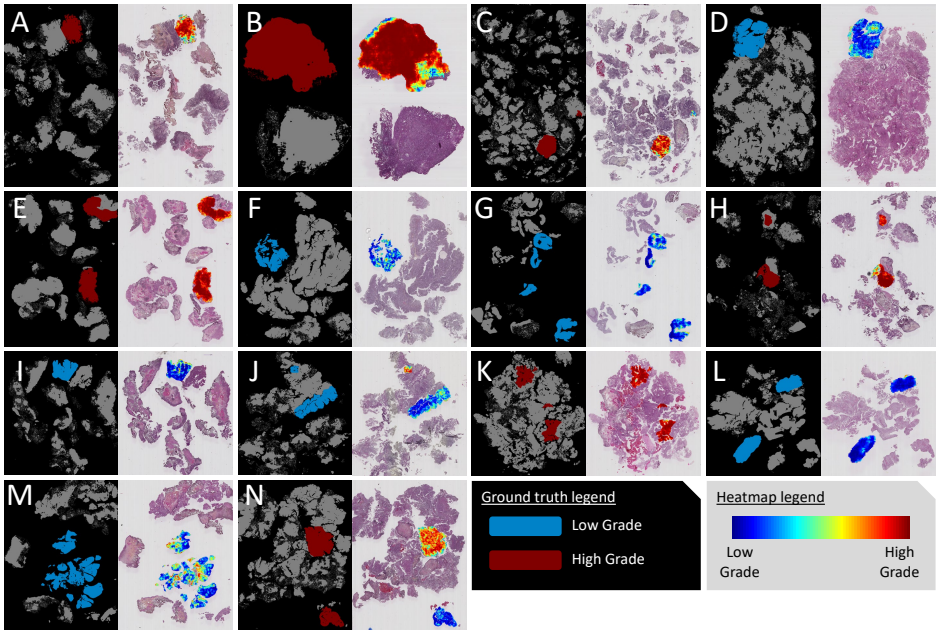


Figure 7.4: Ground truth annotations vs. model prediction. The WSIs with a black background is the ground truth images with low- and high-grade annotations. The WSIs with a grey background has superimposed a heatmap from the same area as the ground truth and highlights the predictions from the $\text{TRI}_{\text{WHO04}}$-model.

7.6 Conclusion

In this paper, we have proposed a $\text{TRI}_{\text{grade}}$ pipeline for automatic grading of urothelial carcinoma slides based on the WHO04 grading system. First, the slide is segmented into the tissue classes (urothelium, stroma, muscle, blood, damaged tissue, and background). Next, tiles are extracted from the urothelium regions at three magnification levels (25x, 100x, and 400x). The three tiles form a triplet, which is fed sequentially to a multiscale CNN-based WHO04 grading model.

The proposed method will generate a tissue segmentation map, helpful for the clinicians to easier find diagnostic relevant regions during an examination. The system will also output a WHO04 grade heatmap, highlighting the most severe urothelium tissue regions, beneficial for the pathologists who can focus their limited per-patient time on the most important regions in the WSI. Finally, the system produces a slide-level WHO04 grade that could potentially be used to prioritize high-grade patients for earlier

examination, as well as suggest the diagnosis to the pathologist.

The system as a whole can be used by clinicians and pathologists to potentially improve their decision-making and further help patients by receiving correct diagnoses and treatment.

Chapter 8

Discussion and conclusion

The following chapter will give an overview of the thesis as a whole. The objectives of the thesis were established in Chapter 1 and will now be discussed in light of what has been presented in Chapters 2-7. First, the sub-objectives will be discussed, followed by the main objective. Then, each subsystem will be discussed; the tissue segmentation, multilevel tile extraction, and diagnosis prediction. Thereafter, a usage scenario is given, showing how the proposed methods could be combined into a clinical setting. Next, some future work is suggested. Finally, the thesis is summarized and concluded.

SO_1 - Create an automated system for distinguishing between the different tissue types present in histological whole-slide images of urothelial carcinoma.

As established in the previous work section in Chapter 1, no system for classifying all tissue classes present in whole-slide images of urothelium carcinoma existed prior to this work. In cooperation with pathologists at Stavanger University Hospital, six classes were chosen to segment in the WSIs. A small annotated dataset (Dataset B) was created together with pathologists, and used to train different models for tissue segmentation. Three papers, presented in Chapter 5, explored different techniques to train models and successfully created a system for the task. The system is capable of creating both heatmaps for individual tissue classes and segmentation maps for all tissue classes in one image.

SO_2 - Explore different approaches for unsupervised and semi-supervised learning techniques to deal with missing annotation data.

With the histological images, we have access to a large amount of image data to train deep learning models. However, annotations are scarce, and due to the need for expert opinions, they are time-consuming and expensive to make.

To mitigate the need for large annotated datasets, several alternative methods were explored. In Paper 1, a large unlabeled dataset was used to train an autoencoder model through unsupervised learning. This step was performed to pre-train a CNN model from scratch and then fine-tuned it to the much smaller labeled dataset. In Paper 2, models pre-trained on the ImageNet dataset were utilized in a technique referred to as domain adaptation. And lastly, in Paper 3, we explored two semi-supervised techniques to train the tissue segmentation models.

SO_3 - Investigate the use of multiscale models, by utilizing several magnification scales.

Pathologists will zoom in and out of the tissue specimen to study the tissue at several magnification levels. To try and mimic this behavior, we implemented multiscale models in Paper 2. Three scales are embedded in the WSIs in the dataset, and all three levels were used. First, each level was used separately in the three MONO models. Then, combinations of two levels were explored in three DI models. Lastly, a TRI model was trained using all three magnification scales.

SO_4 - Create a system that automatically extracts tiles from multilevel whole-slide images.

With such a large dataset and automatically generated tissue maps, a system for extracting tiles from specific tissue classes is needed. Furthermore, with the use of multiscale models, tiles need to be extracted from all three scales at the same area in the WSIs. For this, a system was implemented and presented in Paper 4. The system is parameterized, meaning that a few parameters determine the behavior of the system. Thus, the system is both repeatable and reproducible, and easy to report.

O_1 - Create a system for automated grading of urothelial carcinoma slides.

WSIs of urothelial carcinoma are graded manually by pathologists, but this process suffers from intra- and interobserver variability. A system for grading these WSIs automatically could aid pathologists by providing a second opinion.

A system for grading WSIs was proposed in Paper 5. The system uses the tissue model to extract the urothelium tissue, and the multilevel tile extraction method to generate the diagnostic datasets. Because the multiscale models utilizing all three magnification scales performed the best, the diagnostic model inherits this property and also uses all scales. In addition to providing a slide-level grade for the WSI, a heatmap is also generated, visualizing low- and high-grade carcinoma areas in the WSIs.

8.1 Tissue segmentation

The topic of tissue segmentation was explored quite thoroughly, with three papers devoted to the subject.

In Paper 1, an autoencoder method was used to train unsupervised on a large unlabeled dataset. One of the drawbacks of this method, is the difficulties in training the models, and also the amount of training time. First, the encoder-decoder model is trained on the AE dataset. A large hyperparameter search is necessary to explore most of the options, as it is not clear which parameters to use. In addition, it is not as easy as by looking at the loss between the input and generated output. It is possible to make this loss go towards zero by using a large enough latent vector. However, the goal is not perfect reconstruction, but rather classification, which we need to train a new model on a different dataset to check. As mentioned, training time is also considerable for training both the encoder-decoder and encoder-classifier. The unlabeled dataset, Dataset A in Figure 4.7, is by far the largest dataset of all the datasets used in this thesis. The models are also trained from scratch, resulting in many epochs.

To try and mitigate some of these difficulties, a new approach was taken in Paper 2. Here, pre-trained models were used together with domain adaptation. This reduced the number of necessary epochs and the total training time. In addition, this paper introduced multiscale models, which outperformed the single-scale models.

The best model from Paper 2 was used in Paper 3 and used to explore two semi-supervised learning techniques. Using the model itself to extract a new and larger dataset and retrain the model on both datasets improved the performance.

A significant amount of time was spent composing and implementing techniques for visualization of the tissue classes. Heatmaps were tried out first, which provided the intended information, but did not convey it satisfactorily. The heatmaps were separate images from the WSI, so both the WSI and the heatmap are needed to interpret them. In addition, one heatmap is generated for each tissue class, making it unpractical. In Paper 2, a new tissue segmentation image was proposed, which displayed all tissue classes on one image, and was a far better solution.

8.2 Multilevel tile extraction

The methods presented in Paper 4 for extraction of tiles in multilevel gigapixel images work exceptionally well. The methods are quick, effective, and versatile. It was developed primarily for extracting tiles and visualizing tiles from all levels on the same image. However, after some experimental work, it was concluded that the method also works for post-processing and is suitable for creating heatmaps.

Earlier, more primitive versions of the methods were used throughout Paper 1-3 to extract tiles. However, it was not until the start of work on the diagnosis dataset that the methods were further developed and completed that we looked at the possibility of publishing it.

Only a few parameters are necessary to describe the behavior of the methods. In Paper 5, a small section was included describing the parameters used for tile extraction, demonstrating the simplicity of reporting the method.

8.3 Diagnostic prediction

The whole work of this thesis leads up to the final objective, making an automated diagnosis system for grading urothelial carcinoma slides.

In Paper 1, the generated heatmaps were separate images. Whereas in Paper 5, the improved methods from Paper 4 were used to generate the heatmaps. This means that the heatmaps could be superimposed upon the WSI itself, generating more pleasing and easier to interpreter images.

8.4 Usage scenario

The proposed system in this thesis could potentially be implemented in a clinical setting. A rack of glass slides with prepared specimens is fed into the slide scanner. The slide scanner will automatically scan and save each of the glass slides in the rack to a connected computer, and the saved WSIs are then ready for a pathologist for examination. The proposed system could be installed on the connected computer, and when a new WSI is detected, the proposed system will automatically start analyzing it. This is a background process that can run 24 hours of the day without supervision. When a pathologist is ready to examine the WSI, he or she will now also have access to the segmented tissue map, a WHO04 grade heatmap, and the slide-level WHO04 grade prediction. These are automatically provided to the pathologist as tools to aid them during the examination, potentially making their diagnosis more accurate and speeding up their examination by increasing the efficiency.

It is also possible to take it one step further. Instead of every hospital having its own GPU computer and software connected to the slide scanner, it is possible to have the program running in the cloud. Each hospital can then subscribe to a service where the slides they scan will be processed in the cloud, and the resulting images and predictions will be presented to them in the same way as described above. This has many benefits, like reducing downtime due to failure of equipment, lower maintenance, and also opens up the possibility for hospitals in low-income countries to subscribe to the service, which would otherwise not have the option to buy expensive computers necessary. A drawback, however, is the need for a high-speed internet connection due to the large size of the WSIs.

With the suggested system above, it is possible to use the automatic WHO04 grades to prioritize some patients in front of others. For example, a patient with a prediction of high grade could have a shorter time before a potential recurrence compared to a patient with a prediction of low grade.

8.5 Suggested future work

The work in this thesis covers a lot of different aspects. However, the amount of unexplored items far outweighs the items covered. Therefore, some suggested future work is discussed here.

Deep learning models suffer from the "black box" aspect. That is, when a model makes a prediction, it is unknown to us what made the model come up with that prediction. Being used as a tool in a clinical setting, it would be beneficial if the model could provide some insight into how that decision was reached. For example, attention modules can help to give insight into which parts of a classified image contribute to the predicted class. It could be used as a tool for researchers to identify these regions, or to improve the accuracy and create more robust models. Another aspect is explainable AI (XAI), where the AI models provide interpretability and explainability for its predictions. Amann et al. [2] explore the role of XAI in clinical decision support systems.

At present, the WSI, segmented tissue maps, and WHO04 grade heatmaps are separate images. However, this is not optimal as the user needs to switch between the images. A better solution would be to incorporate the tissue map and grade heatmap as overlays on the WSIs in a GUI software, with a button to toggle the visibility of these on and off.

In the diagnostic model, the urothelium tissue is extracted and used to predict the grade. However, the other classified tissue classes are left untouched. This means that tiles of lower magnification scale (25x, 100x) may cover an area outside the urothelium regions and may include tissue classes like damaged tissue or blood. This is unwanted and may harm the prediction accuracy. By using the tissue map, these classes can be masked out and excluded from any predictions.

The quality of a WSIs can be divided into two parts, tissue and image quality, and was presented in Chapter 4.4. Even though the data material used in this thesis consisted of high-quality slides, if the system should be used on new slides in the future, a quality control system could be used to screen bad WSIs. This was not handled in the thesis but should be taken into consideration in a clinical setting. Several methods for detection of scanning artifacts such as out-of-focus areas have been proposed [63, 87, 109]. As well as open-source tools, such as HistoQC [58], could also be useful for this task.

The stain color varies due to differences in tissue and dye, and is also dependent on laboratory protocols and the slide scanner manufacturer. Methods for stain color normalization have been proposed [48, 107], and could be implemented as a preprocessing step. On the other hand, color augmentation is a technique used during training to make the models more robust against color changes, for example, as proposed by Wagner et al.

[132]. In the work of Tellez et al. [121], both of these techniques were quantified, and they concluded that "*combining color augmentation and color normalization achieves the best performance*".

8.6 Conclusion

The ultimate goal is to develop new tools for pathologists by leveraging digital pathology and digital versions of the tissue samples. This thesis proposes several new tools, including tissue heatmaps, tissue segmentation maps, WHO04 grade heatmaps, and automatically slide-level WHO04 grading. These are all tools that can aid pathologists in becoming more efficient in their work, saving precious time. It could further contribute to the pathologists becoming more accurate in their diagnosis work and help mitigate the problems related to intra- and interobserver variability and inconsistent reproducibility between pathologists. More accurate diagnosis would lead to a more accurate treatment program for the patients, potentially reducing under- or overtreatment. Another potential benefit for the patient is prioritizing high-grade patients based on the slide-level predictions, where the patient who needs correct treatment quickly is diagnosed first and does not have to wait in line.

Another proposed method in this thesis is extraction of tiles in multilevel gigapixel images. It does not directly aid pathologists and patients but could indirectly benefit them by supporting the researchers developing new tools for histological images. The methods were a vital part of this thesis and were used to extract the necessary dataset for the diagnostic models.

Many learning techniques were explored, including unsupervised, semi-supervised, domain adaptation, and supervised learning. The combined results from all these experiments may aid other researchers considering similar approaches.

Paper 1: Multiclass Tissue Classification of Whole-Slide Histological Images using Convolutional Neural Networks

Multiclass Tissue Classification of Whole-Slide Histological Images using Convolutional Neural Networks

Rune Wetteland[1], Kjersti Engan[1], Trygve Eftestøl[1], Vebjørn Kvikstad[2], Emiel A. M. Janssen[2,3]

[1] Department of Electrical Engineering and Computer Science, University of Stavanger, Norway

[2] Department of Pathology, Stavanger University Hospital, Norway

[3] Department of Mathematics and Natural Sciences, University of Stavanger, Norway

Published by SciTePress in the Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods (ICPRAM), 2019.

https://doi.org/10.5220/0007253603200327

Abstract:
Globally there has been an enormous increase in bladder cancer incidents the past decades. Correct prognosis of recurrence and progression is essential to avoid under- or over-treatment of the patient, as well as unnecessary suffering and cost. To diagnose the cancer grade and stage, pathologists study the histological images. However, this is a time-consuming process and reproducibility among pathologists is low. A first stage for an automated diagnosis system can be to identify the diagnostical relevant areas in the histological whole-slide images (WSI), segmenting cell tissue from damaged areas, blood, background, etc. In this work, a method for automatic classification of urothelial carcinoma into six different classes is proposed. The method is based on convolutional neural networks (CNN), firstly trained unsupervised using unlabelled images by utilising an autoencoder (AE). A smaller set of labelled images are used to train the final fully-connected layers from the low dimensional latent vector of the AE, providing an output as a probability score for each of the six classes, suitable for automatically defining regions of interests in WSI. For evaluation, each tile is classified as the class with the highest probability score. The model achieved an average F1-score of 93.4% over all six classes.

9.1 Introduction

Globally, bladder cancer resulted in 123 400 deaths in 1990, and in 2010 this number was 170 700 which is an increase of 38.3% taking population growth into consideration [76]. The majority of bladder cancer incidents are urothelial carcinoma with a representation as high as 90% in some regions [37]. For patients diagnosed with bladder cancer, 50-70% will experience one or more recurrences, and 10-30% will have disease progression to a higher stage [81]. Patient treatment, follow-up and calculating the risk of recurrence and disease progression depend primarily on the histological grade and stage of cancer. Correct prognosis of recurrence and progression is essential to avoid under- or over-treatment of the patient, as well as unnecessary suffering and cost.

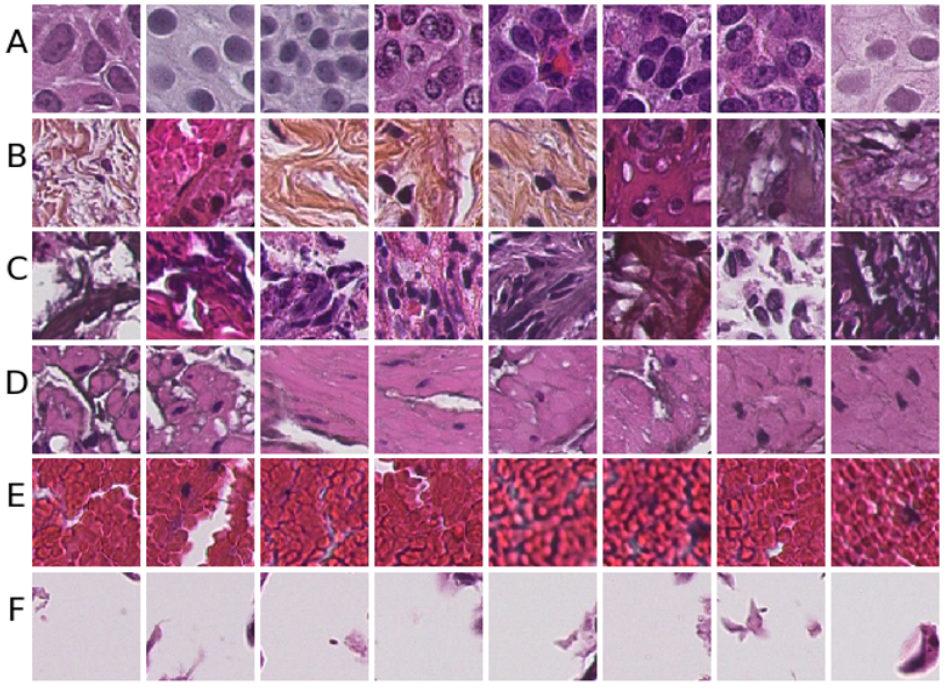


Figure 9.1: Example tiles from each class. A) Urothelium, B) Stroma, C) Damaged tissue, D) Muscle tissue, E) Blood, and F) Background.

With the introduction of digital pathology, some computer-aided tools to assist pathologists have been introduced, but still the assessment of histopathological images to diagnose, grade and stage cancer is mainly

done manually. This is a time-consuming process and reproducibility among pathologists is in some cases low, for example within the prognostic classification of urinary bladder cancer. Automatic extraction of the relevant areas in large whole-slide images (WSI) would be an important first step where the results could be used in automated diagnostic and prognostic classification tools.

During the biopsy, parts of the tissue get both physical- and heating-damage, and thus can not be used as relevant diagnostic information. The WSI also contains stroma- and muscle-tissue as well as areas of blood. In this paper we consider the task of automatic classification of tiles in WSI into the six different classes; urothelium, stroma, damaged tissue, muscle, blood and background. Examples from each class are shown in Figure 9.1. The system uses the automatic classification tool to produce heat maps from the model's output. Such heat maps can provide useful information to help the pathologist to focus on the diagnostic important part of the large WSI during visual inspection. In addition, the heat maps are also suitable as input for automatic region of interest (ROI) extraction of relevant areas in the WSI, which can further be used in automated diagnostic and prognostic classification tools.

9.1.1 Previous work

In recent literature, some methods for automatic tissue classification have been suggested. However, most previous works have focused on classifying only two classes, a binary problem set to differentiate between cancer-patches and non-cancer patches.

Recent literature shows good results for binary tissue classification using convolutional neural networks (CNN). Wang et al. [133] won both competitions of the Camelyon16 grand challenge for automated detection of metastatic breast cancer in WSI. As part of their model, GoogLeNet was utilised to do patch classification. The model was trained to discriminate between positive and negative patches and achieved an accuracy of 98.4%.

Some attempts of multiclass tissue classification can be found in recent years. Araújo et al. classified patches of breast cancer into four classes using convolutional neural networks [5]. The best patch-wise accuracy for four classes was 66.7%. When the task was simplified as a two-classes problem, non-carcinoma vs carcinoma, the accuracy was improved to 77.6%. The work of Kather et al. [62] uses a combination of several hand-crafted feature

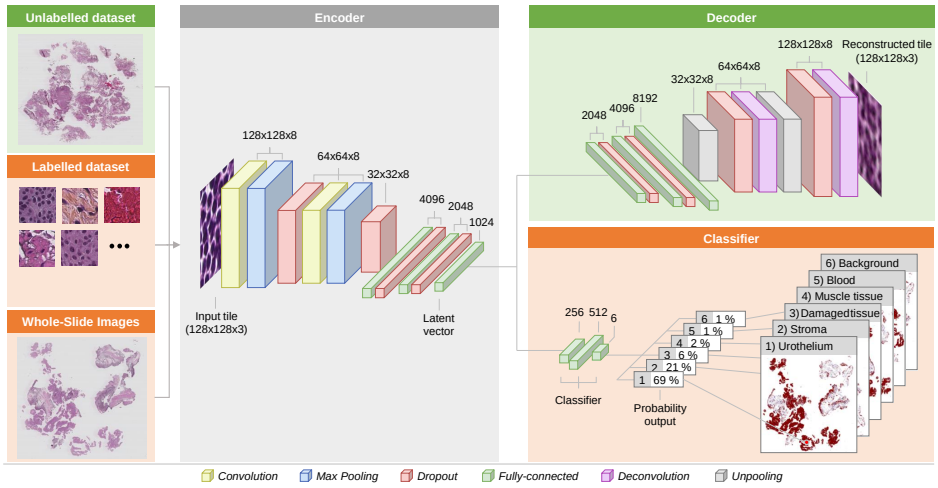


Figure 9.2: Overview of the CNN-model. First, the unlabelled dataset is used to train the encoder-decoder model. Then the labelled dataset is used to train the encoder-classifier model. Finally, the trained encoder-classifier model are used to classify new WSI into probability maps. These probability maps are further postprocessed to produce the heat maps.

methods to classify different types of tissue in colorectal cancer, performing tests on both a two-class and eight-class problem. They achieved the best result on the two-class problem with a tumour-stroma separation accuracy of 98.6%, while the multiclass problem achieved an accuracy of 87.4%.

To the author's knowledge, there are no published results on multiclass classification on WSI of bladder cancer.

Some few and recent work on ROI detection can be found. ROI detection has been done by multi-scale real-time coarse-to-fine topology preserving segmentation (CTFTPS) by utilising superpixel clustering technique [69, 145]. A RAPID (Regular and Adaptive Prediction-Induced Detection) segmentation method for ROI detection in large WSI is presented by Sulimowicz and Ahmad [118] while using the multi-scale CTFTPS technique as a baseline. An SVM was utilised to classify the detected regions as ROI vs non-ROI. For this task, the classifier achieved an F1-score of 89.8% for the RAPID method, and 91.2% for the optimised multi-scale CTFTPS method.

Deep CNN has shown to provide state of the art results in many computer vision tasks in recent years [66] and has also found its way into medical image assessment tasks. In this work, a method for automatic classification

of WSI from urothelial carcinoma into six different classes is proposed. The method is based on CNN, firstly trained unsupervised, using large unlabelled image sets by utilising an autoencoder (AE). A set of labelled images are used to train the final fully-connected layers from the low dimensional latent vector of the AE, providing an output as a probability score for each of the six classes, suitable for automatically defining ROI in WSI. A visualisation of the system is depicted in Figure 9.2.

The novelty of the work lies both in the specific application of urinary bladder WSI and in the method development, more specifically in a combination of using CNN, learned in a semi-supervised way, for the application of automatic region of interest extraction in WSI by *multiclass* tissue classification, tested on urinary bladder cancer.

9.2 Data material

The data material used in this paper consists of histopathological images from patients with primary bladder cancer, collected in the period 2002-2011 at the University Hospital of Stavanger in Norway. The biopsies are formalin fixed and paraffin embedded, 4 µm slides are cut and stained with Hematoxylin Eosin Saffron (HES). All slides are diagnosed and graded according to WHO73 and WHO04, cancer stage (Tis, Ta or T1) and follow-up data on recurrence and disease progression are recorded.

The slides are then scanned using a Leica SCN400 histological slide scanner to produce a digital histological image. The images are in Leicas data format called SCN and to be able to process these images the Vips library [84] has been used, which is specially designed for image processing of large images.

9.3 Proposed method

An overview of the proposed method can be seen in Figure 9.2. The different parts will be explained in this section.

9.3.1 Preprocessing

Each WSI is sliced into smaller non-overlapping tiles of size 128 $\times$ 128 pixels, extracted at 400x magnification level. The background takes up as

much as 70-80% of the WSI and is detected and discarded automatically by computing the histogram of the tile and setting a fixed threshold value. This removes tiles consisting of grey background, however, if the background tile contains small parts of debris, tissue or similar it is not discarded. Examples of tiles belonging to this class are illustrated in Figure 9.1-F.

The histological images are split into three datasets. First, an unlabelled dataset is created in the manner explained above where the extracted tiles have no label associated with it. In total 48 WSI all from different patients were preprocessed resulting in 7 130 527 unlabelled tiles after the pure background tiles are excluded. This set, called *train-ae*, is utilised as training data for the AE-model.

Secondly, a labelled training dataset is created. A pathologist has manually annotated carefully selected regions in the WSI. The tiles in the regions are preprocessed by evaluating the histogram to be sure not to include background or boundaries and given a label corresponding to its class. The number of patients and tiles produced are listed as *train-set1* in Table 9.1.

Lastly, a labelled test set is created to assess the performance of the classifier. The set is created in the same manner as the labelled training set, but on separate WSI which has not been used in either the unlabelled or labelled datasets to avoid cross-contamination between training and test data. The dataset is listed as *test-set* in Table 9.1.

The texture of urothelium tissue will change for the different cancer grades, and thus it is vital to include a wide variety of samples for this class. The other five classes, however, will not change as a function of cancer grade and may include fewer samples. Another issue is that the occurrence of some classes is more sparse in the WSI, making it difficult to extract a large amount of it. A disadvantage of these two issues is a significant deviation in the number of samples in two of the classes, stroma and muscle tissue, as seen in *train-set1* in Table 9.1.

To compensate for the class-imbalance in *train-set1*, data augmentation techniques have been utilised. Tiles in the muscle and stroma class are extracted with 50% overlap, to produce more data from the same regions. These extracted tiles are further augmented by randomly flipping and rotating them to create new data. These techniques result in a more balanced dataset, which is listed in Table 9.1 as *train-set2*. This dataset is used to train the classifier in the presented experiments. The augmentation techniques were not performed on the *test-set*, resulting in an unbalanced test set. In this case, accuracy as a performance metric could be misleading. Instead, precision, recall and F1-score are used to evaluate the performance.

Table 9.1: The resulting labelled datasets after preprocessing. Results show the total number of tiles extracted for each class, and the number of WSI used are shown in parenthese.

	Train-set1	Train-set2	Test-set
Urothelium	25 635 (25)	25 635 (25)	3 612 (3)
Stroma	4 329 (4)	25 974 (4)	505 (1)
Damaged	30 714 (8)	30 714 (8)	2 679 (1)
Muscle	2 002 (3)	23 949 (3)	475 (1)
Blood	19 071 (4)	19 071 (4)	692 (1)
Background	20 000 (2)	20 000 (2)	500 (1)

9.3.2 CNN-Model

The system consists of an autoencoder model which is trained on the unlabelled dataset *train-ae*. The autoencoder consists of two main parts; the encoder and the decoder. The encoder will transform the input tile into a latent vector of much lower dimension. A small latent space is chosen which will force the network to extract the essential features of the image and preserve these in the vector. The decoder will use the features stored in the latent vector and reconstruct the input. During training, the network compares the reduced mean of the squared difference between the input image and reconstructed output image as given by the loss function $\sum(input - output)^2$. The AE function is described in details in [8]. The encoder consists of two convolutional-, two max-pooling- and four dropout-layers, as well as three fully-connected layers as seen in Figure 9.2. The decoder consists of the same layers, but in reverse order and uses unpooling and deconvolutional layers instead.

After training, the encoder has learned to extract the features of the input tile, which are now stored in the latent vector. To do classification, the decoder part is discarded and exchanged with a classifier. The classifier consists of three fully-connected layers connected to the output of the encoder. This encoder-classifier model constitutes the proposed CNN-model and is trained on the labelled training dataset *train-set2* and evaluated on the *test-set*.

For initialisation of the system, the bias is set to zero, and the weights are taken from a truncated normal distribution. The convolutional layers

use a filter kernel of 3×3 and a stride of 1, whereas the max-pooling layers use a filter kernel of 2×2 with a stride of 2. The number of feature maps is used to control the size of the latent vector space and is experimented on as described in Section 9.4. The parameters of the network are optimised using the Adam optimiser with a mini-batch of size 128. For the activation function between layers, the rectified linear unit (ReLU) activation function is used. For the last layer, the softmax activation function is utilised. This will output a true probability distribution, meaning each output lays in the interval 0 to 1 and all outputs combined sums up to one. Dropout is a technique where randomly selected nodes are set to zero during training to provide regularisation to the network. The portion of nodes set to zero is specified by the dropout rate as a percentage. During evaluation of the network, dropout is disabled.

The histological images are in Leicas data format called SCN and to be able to process these images the Vips library [84] has been used. This is a library specially designed for image processing of large images. The model is written in Python 3.5 using the Tensorflow 1.7 machine learning library [1]. For evaluation of the model, the Scikit-learn metric package [98] is used which computes precision, recall and F1-score of each class in addition to an average total score.

The model is used to predict the class of each tile in a WSI. The probability for each class provided by the model can be rearranged as probability maps, one for each class, and will visualise the location in the histological image where each class is present. An overview of this process is presented in Figure 9.2.

9.4 Experiments and results

Two experiments were conducted, the first to find the best combination of architecture and hyperparameters and the second to verify its performance and use the final model on WSI.

9.4.1 Experiment 1: Architecture and hyperparameters

To find a suitable architecture and appropriate hyperparameters, a large grid search was conducted. To reduce both computational time and search space, a preliminary search was set up with some limitations. A reduced

version of the *train-ae* dataset was used to decrease the processing time, and each model was only trained for 50 epochs.

The encoder-decoder model was tested with two different sizes of the latent vector, which was altered by changing the number of feature maps in the convolutional layers. Latent vectors of size 512 and 1 024 were tested. A learning rate of 10^{-3} and 10^{-4} was tested as well as dropout rates of 0%, 10% and 20%. Each of these combinations was tested on network configuration consisting of two, four and six convolutional layers in the autoencoder.

In the encoder-classifier model, the classifier consists of three dense layers. The first layer after the encoder was tested with 256, 512 and 1 024 neurons, and the second layer with 128, 256 and 512 neurons. The number of neurons in the output layer is bounded to the number of classes. This results in 9 different configurations for the classifier layers. Each of these configurations was tested with a learning rate of 10^{-3}, 10^{-4} and 10^{-5}. There are no dropout layers in the classifier itself, but changing the dropout rate will affect how the encoder codes the input tile into the latent vector. The encoder-classifier was therefore also tested with the same dropout rates as above. The model was tested both with and without freezing the pre-trained encoder-layers to see how it affected the result.

The prediction accuracy on the *test-set* was used to compare the performance of the different hyperparameter combinations. Hyperparameters that showed poor performance on several models were excluded to narrow down the search space.

The experiments showed an overall best result using an encoder-decoder structure with two convolutional layers with a latent vector of 1 024 neurons trained with 10^{-4} learning rate and 10% dropout rate. The results further showed best performance while not freezing the encoder part of the encoder-classifier model. A classifier with 256 neurons in the first layer and 512 in the second layer was favourable, trained using a learning rate of 10^{-5} and 10% dropout rate. These hyperparameters and settings will be used as the resulting model of this experiment. The model is depicted in Figure 9.2.

9.4.2 Experiment 2: Training, testing and using the resulting model

The resulting architecture after the first experiment was trained once more, this time on the full dataset. First, the autoencoder was trained on the unlabelled dataset *train-ae* for 100 epochs, then the encoder-classifier was

fine-tuned on the augmented labelled dataset *train-set2* for another 600 epochs. Since experiment 1 showed best results when the encoder was not frozen during fine-tuning, both the encoder and classifier was trained during this step. Evaluation using the Scikit-learn metric package on the *test-set* was performed every 5^{th} epoch. The model achieved the best result after 540 epochs of training with an average F1-score of 93.4% over all six classes. The precision, recall and F1-score of each class is shown in Table 9.2.

Table 9.2: Detailed classification results from the model trained using 10% dropout rate.

Class	Precision	Recall	F1-Score
Urothelium	0.924	0.952	0.938
Stroma	0.897	0.929	0.913
Damaged	0.925	0.927	0.926
Muscle	0.980	0.714	0.826
Blood	0.996	0.991	0.994
Background	0.990	0.988	0.989
Average total	0.936	0.935	0.934

The overall results in Table 2 are good. However, there are some observations.

In *train-set2*, which is used to train the classifier, the classes of blood and background have the fewest number of samples. However, these are the classes which perform best. This is probably because these classes have the least within-class variance, e.g. most of the tiles have a similar visual appearance.

Urothelium and damaged tissue both perform well, even though these classes have a substantial visual variance in the form of colour and texture in the tiles. The dataset for these classes contains the most number of patients (25 and 8 patients, respectively), and therefore contains the most diverse samples in the dataset, contributing to the good results.

The precision of stroma and recall of muscle is not performing as good as the rest. The dataset for these classes contains few patients and are also the two classes which needed augmentation due to small amounts of

available data. The low recall of muscle tissue indicates that a large proportion of the muscle tiles are misclassified as other classes, most probably urothelium, stroma and damaged tissue (due to the high precision of blood and background, these are not likely to include many misclassified tiles). It is important to note that the muscle class achieves a very good precision score, and stroma has an acceptable good recall score.

9.4.3 Heat maps

The resulting model was utilised to classify entire whole-slide images. Each tile in the WSI was classified and the percentage for each class recorded. These were then combined to create the probability maps. These maps were then post-processed in MATLAB by applying a Gaussian filter kernel with a standard deviation of $\sigma = 0.6$ to smooth the images. After filtering, a thresholding operation was performed on the image with a limit of 0.8, setting all predictions below this threshold to zero. This ensures that only predictions of 0.8 or higher are visible in the final heat maps.

Figure 9.3 shows three example WSI with their corresponding heat maps. By visual inspection performed by pathologists, this is considered to look very promising. However, a quantitative measure for the WSI ROI extraction is lacking since we do not have complete WSI manually labelled into the six classes at the current time.

9.5 Conclusion

This paper proposes a method for automatic classification of tile-segments of histopathological WSI of urinary bladder cancer into six different classes using a CNN-based model. An encoder-decoder structure is trained on a large set of unlabelled data. After training, the encoder part of the autoencoder acts as a feature extractor making low dimensional latent vectors. An encoder-classifier structure is then fine-tuned on a set of labelled tiles. The finished model is able to classify input tiles from the WSI into the classes urothelium, stroma, damaged tissue, muscle, blood and background. The best model achieved an average F1-score of 93.4% over all the six classes, an overall good result. However, future work will include an effort to improve the classifier. Other methods such as a multiscale approach are considered.

The model is further used to classify entire WSI to produce heat maps, which visualises each of the classes and their location in the image. These maps can provide useful information to the pathologist during visual inspection. Future work consists of using the above model as an ROI extractor of relevant tissue in the WSI to make a dataset suitable as training data for a diagnostic and prognostic classification model.

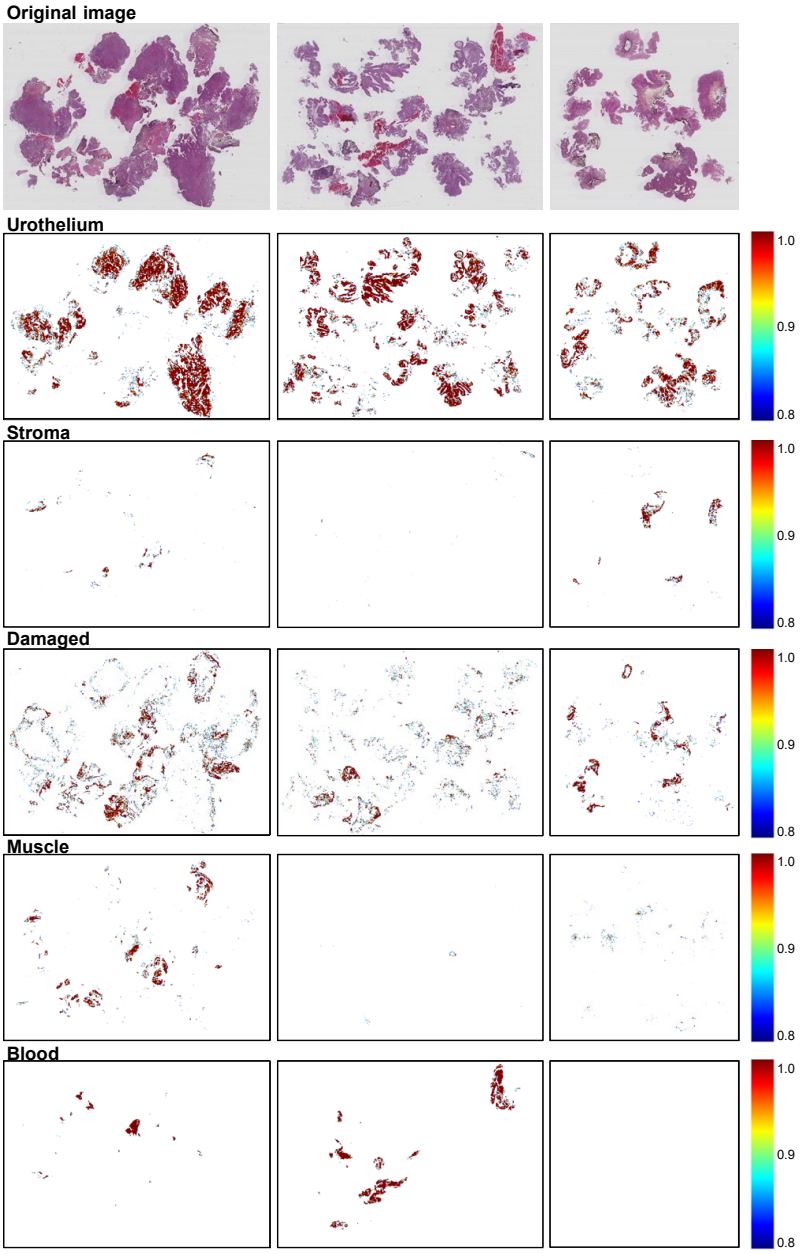


Figure 9.3: The original WSI together with the corresponding heat maps. The scale in the rightmost column shows the confidence level given by the model. The background heat maps are performing very good, but has been omitted from the heat map visualisation since it is just removing the borders between background and tissue. The heat maps have been smoothed with a Gaussian filter and thresholded to only contain predictions of 0.8 and higher.

Paper 2:
A Multiscale Approach for Whole-Slide Image Segmentation of Five Tissue Classes in Urothelial Carcinoma Slides

A Multiscale Approach for Whole-Slide Image Segmentation of Five Tissue Classes in Urothelial Carcinoma Slides

Rune Wetteland[1], Kjersti Engan[1], Trygve Eftestøl[1], Vebjørn Kvikstad[2], Emiel A. M. Janssen[2,3]

[1] Department of Electrical Engineering and Computer Science, University of Stavanger, Norway

[2] Department of Pathology, Stavanger University Hospital, Norway

[3] Department of Mathematics and Natural Sciences, University of Stavanger, Norway

Published in the Journal of Technology in Cancer Research and Treatment (TCRT), 2020.

https://doi.org/10.1177/1533033820946787

Abstract:
In pathology labs worldwide, we see an increasing number of tissue samples that need to be assessed without the same increase in the number of pathologists. Computational pathology, where digital scans of histological samples called whole-slide images (WSI) are processed by computational tools, can be of help for the pathologists and is gaining research interests. Most research effort has been given to classify slides as being cancerous or not, localization of cancerous regions, and to the "big-four" in cancer: breast, lung, prostate, and bowel. Urothelial carcinoma, the most common form of bladder cancer, is expensive to follow up due to a high risk of recurrence, and grading systems have a high degree of inter- and intra-observer variability. The tissue samples of urothelial carcinoma contain a mixture of damaged tissue, blood, stroma, muscle, and urothelium, where it is mainly muscle and urothelium that is diagnostically relevant. A coarse segmentation of these tissue types would be useful to i) guide pathologists to the diagnostic relevant areas of the WSI, and ii) use as input in a computer-aided diagnostic (CAD) system. However, little work has been done on segmenting tissue types in WSIs, and on computational pathology for urothelial carcinoma in particular. In this work, we are using convolutional neural networks (CNN) for multiscale tile-wise classification and coarse segmentation, including both context and detail, by using three magnification levels: 25x, 100x, and 400x. 28 models were trained on weakly labeled data from 32 WSIs, where the best model got an F1-score of 96.5% across six classes. The multiscale models were consistently better than the single-scale models, demonstrating the benefit of combining multiple scales. No tissue-class ground-truth for complete WSIs exist, but the best models were used to segment seven unseen WSIs where the results were manually inspected by a pathologist and are considered as very promising.

10.1 Introduction

Worldwide, 549 393 new cases of bladder cancer were diagnosed in 2018, in addition there were 199 922 deaths due to the disease. This makes bladder cancer the 10^{th} most common type of cancer in the world [15]. Men are overrepresented, with approximately 75% of the cases [4]. The most common type of bladder cancer is urothelial carcinoma, with over 90% of the cases [37]. Of the patients diagnosed with bladder cancer, 50% to 70% will experience recurrence, and 10% to 30% will advance to a higher disease stage [81].

Treatment and follow up of urothelial carcinoma are primarily based upon histological grade and stage, evaluated manually by an expert pathologist studying the histological images of the tumor using the latest WHO16 classification system [7]. Correct grade and stage are essential to avoid over- or under-treatment, and thereby unnecessary suffering for the patient. For most pathology departments, evaluation of histological images is still performed through a microscope, a time-consuming process, not always reproducible [82]. Digital pathology has been introduced to improve diagnostic accuracy, and certain computer-aided diagnostic (CAD) tools are in use for other diseases. However, such tools are currently not in use for the assessment of urothelial carcinoma and could potentially be of great value to patients and clinicians.

Non-muscle invasive bladder cancer is usually treated with transurethral resection of the tumor. The removed tissue contains both atypical urothelium from the tumor as well as stroma, but can also contain smooth muscle from the bladder wall, normal urothelium from surrounding mucosa and blood. During the procedure, parts of the tissue can get damaged, for example in terms of heating damage induced by laser or electrically heated wire loop. Areas on the whole-slide images (WSI) with blood and damaged tissue will not be suitable for extracting diagnostic and prognostic information, and a pathologist will discard such regions on inspection. CAD systems processing WSI must be able to identify trustworthy interesting areas of resected tissue, but also identify damaged areas and regions that should be excluded from further analyses.

This paper proposes an automatic method for classifying WSI tiles from urothelial carcinoma cases into the following categories: urothelium, stroma, muscle, damaged tissue, blood, and background, utilizing different magnification scales. Examples from each class are shown in Figure 10.1. The output of such a system can be used as a guide for pathologists, providing

a quick visualization of where the different tissue types can be found. To the best of the author's knowledge, a system for segmenting urothelial carcinoma WSIs into each tissue class does not exist. For determination of stage, pathologist wants to identify if muscle tissue is present or absent in the WSI and whether the tumor has infiltrated it. As muscle tissue is often sparse in the WSI, it can be time-consuming to get a full overview of its locations. However, with the help of segmented tissue images, it can be verified in a short amount of time. In the future, training data for a CAD system will be created by utilizing the best model developed through this paper by extracting diagnostic relevant features from the appropriate and relevant regions in the WSI. As this problem is not strictly dependent on classifying all six tissue classes, a binary approach is also experimented with in this paper classifying only urothelium vs. non-urothelium tissue to see if an increase in urothelium extraction can be achieved.

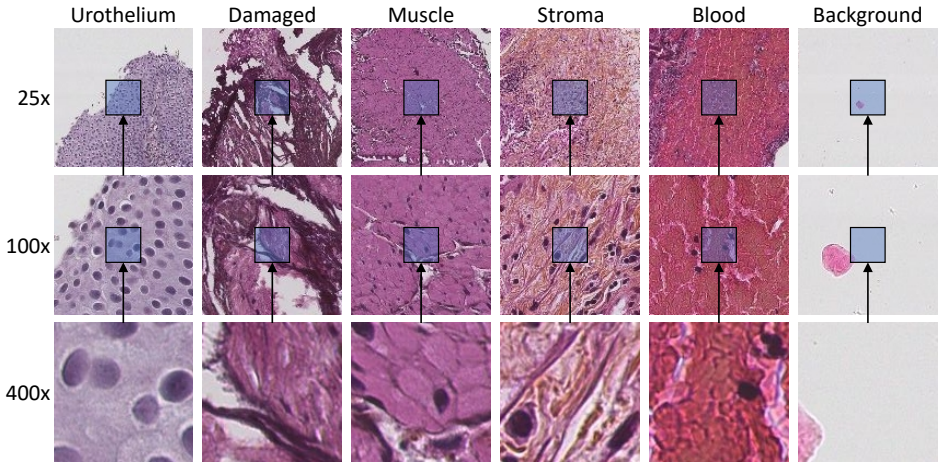


Figure 10.1: Example tiles of each class extracted at three magnification scales. Tiles at each scale are extracted from the same center pixel. The magnification scale is increased by a factor of four in each step, resulting in the tile covering 16 times as much area, even though they have the same size of 128 × 128 pixels.

Tile-based classification of WSI has been done earlier [5]. However, by only classifying a single tile, it leaves out information from the surrounding area. Moreover, WSI viewed on different magnification scale identifies different information. During an examination, a pathologist will integrate information across several magnification levels before reaching a final decision. Low magnification (25x) will show global context information such as papillary architecture, outline, and the border of the tissue, as well as color

and texture. Nuclear polarity can be evaluated in the mid magnification (100x), while high magnification (400x) will reveal cytological features like cell size and shape, mitosis, as well as cell nucleus characteristics as contour, size and colorization intensity, and distribution.

The proposed method combines global context information found at lower magnifications (25x, 100x) with local information found at the highest magnification (400x) using deep neural networks to extract features from the different scales, thereafter concatenating the features feeding the last classifier layers of the network. Different neural network models were tested which utilized different combinations of the scales.

10.1.1 Related work

It is not possible to feed an entire gigapixel WSI into a deep neural network, and a practical solution to this is to divide WSI into tiles and feed the tiles sequentially to the deep neural network. There are primarily two methods for semantic segmentation within medical applications. The first, which utilizes models capable of providing pixel-wise classifications, can output segmentations with high resolution. These networks are usually based on the fully convolutional networks (FCN) introduced by Long et al. in [75]. Popular models are the U-net model by Ronneberger [104], and variants of this [68, 70]. As these networks can detect small details, they are often used in cell and nuclei segmentation [46, 131], but also on tumor segmentation tasks [151]. The downside, however, is the need for pixel-wise ground-truth annotation for supervised learning, which is difficult and time-consuming to generate, especially in many medical applications. These networks are typically trained and tested on small example-patches from WSIs, since no dataset with a pixel-wise annotation of cells and tissue types on full WSI exist.

The second approach is based on tile-wise classification, where the models output a class label for each tile. This results in a coarser segmentation with the resolution of the tile size, and thus are more often seen for classification tasks rather than segmentation tasks. Nevertheless, it has been used in tumor segmentation methods [19, 49, 55, 56, 142]. As every pixel within the tile belongs to the same class, the tile-based ground-truth annotation process is significantly simplified for classification and localization of regions within histological images.

A combination of both tile-wise and pixel-wise classification has been seen for segmentation of WSI by Guo et al. [47]. Firstly, a tile-based

prediction using Inception-V3 gives a coarse segmentation of the WSI, followed by a pixel-wise classification of only the tumor tiles for refined segmentation of those areas. This approach can speed up the segmentation process relative to a pixel-wise segmentation of the entire slide; however, the need for pixel-wise ground-truth in all region of interests is still a significant challenge.

A pathologist studying a slide would typically zoom in and out, looking at both details and context. To similarly include these features in an artificial intelligence (AI) model, some multiscale approaches have been suggested. Models are trained with multiple input tiles, either taken from different magnification scales or taken from the same scale but with varying sizes to accommodate for a larger field of view. In the work of Sirinukunwattana et al. [114], the author has performed a systematic comparison between five single-scale and five multiscale architectures, tested on four classes of prostate cancer and four classes of breast cancer. Both tiles extracted at different magnification levels, as well as tiles of various sizes, were tested; and the result supports the claim that incorporating a broader visual context improves the outcomes. Another multiscale approach was used by Vu et al. [131], which created a network named multiscale deep residual aggregation network (MDRAN). First, a tile is extracted from the WSI at 200x magnification, and then resized to x0.5 and x2 the original size. The three scales (0.5x, 1x, 2x) were then aggregated in the model and used to accurately segment nuclei of non-small cell lung cancer (NSCLC). Since the models uses multiple inputs, the architectures often become more complex, and the total number of parameters within the models also goes up. This affects both the training and inference time of the models.

Most previous work on WSI classification is targeted on segmenting cancerous vs. non-cancerous areas of the WSI, and often the non-cancerous class may include several tissue classes. E.g., the work just mentioned by Vu et al. [131] also performed WSI classification of NSCLC into three classes: NSCLC adeno (LUAD), NSCLC squamous cell (LUSC) and non-diagnostic (ND). The ND regions, in this case, consisted of fat, lymphocytes, blood vessels, red blood cells, normal stroma, cartilage, and necrosis without any attempt to separate these classes. Sometimes, however, there can be useful information in stroma, muscle, or other non-cancerous tissue types as well. There are some very few reported works on segmenting various tissue types. In [71], Li et al. propose a model with dual inputs trained to segment WSI from the ICIAR2018 breast cancer dataset into normal, benign, situ, and invasive regions. Also, a transfer learning model with

multiple inputs was explored by Wang et al. [134] to segment histological images of inflammatory bowel disease (IBD) into the four categories: muscle regions, messy regions, messy + muscle regions and background. Kather et al. [62] used a deep learning model to classify tiles from colorectal cancer into eight different classes of tissue: tumor epithelium, simple stroma, complex stroma, immune cell conglomerates, debris and mucus, mucosal glands, adipose tissue, and background.

Relatively little work is aimed at segmentation of bladder cancer WSIs. In the work of Xu et al. [142], a method for predicting low or high tumor mutational burden (TMB) in bladder cancer patients was investigated. As a preprocessing step, a tile-wise tumor vs. non-tumor classifier was used to segment out the tumor regions from the surrounding tissue. An SVM classifier was then used to predict the patient's TMB state using extracted histological image features from the tumor regions. A similar approach was used by Zhang et al. [151], where a U-net like network was used to predict each pixel into tumor or non-tumor as a preprocessing step before using another neural network for predicting the slide level diagnosis. As urinary bladder tumors are removed using a laser, burnt and damaged tissue is often present at the WSI. Muscle, stroma, and blood will also be part of the removed tissue and visible in the WSIs. But no effort is aimed at identifying these regions, even though they may contain valuable information for a pathologist.

The recent research efforts show promising results utilizing deep neural networks in different configurations for classifying and localizing cancerous areas. However, most effort is made on the "big four" in cancer (i.e., breast, lung, prostate, and bowel), performed on some publicly available datasets. Still, there is relatively little work done on other cancer types, on multiclass classification, on tissue-type classification, and segmentation/heat maps of full WSI.

10.1.2 Aims and contributions

In Wetteland et al. [138], we presented a method based on convolutional neural networks (CNN) for classifying tiles of urothelial carcinoma WSI into the six classes shown in Figure 10.1. The model utilized the autoencoder architecture and was first pre-trained on a large unlabeled dataset, and afterward fine-tuned on an annotated dataset. The models did not include any context, as both the unlabeled and labeled dataset was extracted at the full image resolution of 400x magnification.

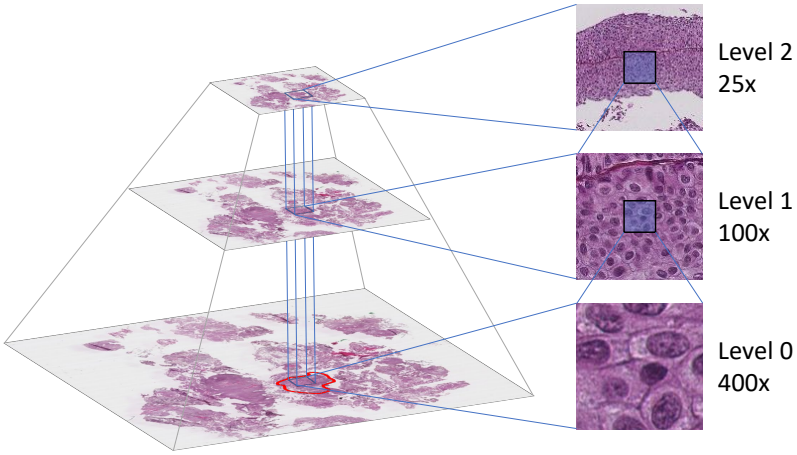


Figure 10.2: The WSI is stored in a pyramidal file format, including several down-sampled versions of the base image. The annotated region (marked with red at level 0) determines which tiles to extract. Tiles are then extracted at the desired location from all three levels.

The main contribution of the current paper is to combine histological images from different magnification scales into the model, giving the model access to a greater field of view and more context of the surrounding tissue. The resulting models are also used to generate segmented images of all the tissue classes within bladder cancer WSIs. An extensive number of experiments are conducted to find the best combination of inputs and magnification levels for the given task. The method utilizes the pyramidical image file format to extract tiles from existing down-sampled versions already present in the file, excluding any up- or down-sampling, limiting the number of necessary computational operations. Transfer learning is incorporated by building on the VGG16 network rather than the autoencoder model. To summarize, this paper proposes an automatic multiscale system, merging inputs of 25x, 100x, and 400x magnification, based on a CNN for classification of whole-slide histological images into six classes.

A preliminary study of this work was published by Wetteland et al. as an abstract [139]. Here we present much more comprehensive experimental work and a description of the method.

10.2 Materials and methods

First, the data material will be introduced and explain how the datasets are prepared. Afterward, the proposed system for tissue segmentation is presented. Then the structure of the model is described, and finally, the training procedure and model selection is explained.

10.2.1 Data material

The data material consists of digital whole-slide images from patients diagnosed with primary papillary urothelial carcinoma, collected at the University Hospital of Stavanger, Norway, in the period 2002-2011. The biopsies are formalin-fixed and paraffin-embedded, from which 4 µm slices are cut and stained with Hematoxylin Eosin Saffron (HES).

The prepared tissue samples are scanned at 400x magnification using the Leica SCN400 slide scanner, producing image files in Leica's SCN file format. The images are stored as a pyramidal tiled image with several down-sampled versions of the base image in the same file to accommodate for rapid zooming. Each level in the file is down-sampled by a factor of 4 from the previous level. Figure 10.2 shows an example of a pyramidal histological image with three levels. The Vips library [84] is capable of extracting the base image as well as the down-sampled versions, making it easy to extract the dataset at each resolution.

Two datasets were collected from the described data material, referred to as the CV dataset and the inference dataset, both are described below.

CV dataset. An expert pathologist carefully annotated selected regions in the WSI, where each region includes one of the six classes. A total of 239 regions belonging to the five foreground classes was annotated in WSI from 32 unique patients. The background regions were extracted from seven randomly selected patients.

The annotated regions contain tight corners and narrow passages to accommodate the shape of the tissue regions in the WSI. When extracting tiles from the WSI, a grid of non-overlapping tiles was superimposed upon the annotated region at 400x magnification level. The tiles in the grid which lie outside of the region are regarded as invalid and will not be used, whereas tiles within the region are valid. By shifting the grid in the X- and Y- direction, more or fewer tiles become valid. To maximize the number of

valid tiles, an automatic search algorithm was developed. The algorithm checks the number of valid tiles for all possible positions of the grid. The grid location with the highest number of valid tiles was used to extract the dataset from that region. This search was performed individually for each region.

Tile sizes of 64 × 64, 128 × 128, and 256 × 256 pixels were tested when extracting tiles with the automatic program. Using a tile size of 64 × 64 extracted the most extensive dataset, but the size may be too small as each tile contain little context information. With a tile size of 256 × 256, the extracted dataset became very small, especially for the stroma and muscle class. A tile size of 128 × 128 was thus chosen as a trade-off between the other two sizes. When a tile is saved from the region, the corresponding tiles from 25x and 100x magnification were also extracted in such a manner that the center pixel is the same in all three magnification levels, as can be seen in the right-half of figure 10.2.

Table 10.1: The resulting CV dataset is listed in the table with the total number of tiles extracted for each class. The number of tiles refers only to tiles extracted at 400x magnification. For the DI- and TRI-CNN models, the numbers need to be multiplied by two and three, respectively. Classes marked with an asterisk shows the number of tiles after augmentation.

Class	Tiles	Patients
Urothelium	29 728	28
Damaged	33 607	9
Stroma*	9 750	5
Blood	19 832	5
Muscle*	19 932	4
Background	27 012	7

The extracted 400x magnification tiles are ensured to stay within the region border. However, by keeping the tile size the same, the lower magnification (25x, 100x) tiles will have a wider field of view, allowing for more context of the surrounding tissue to be included. Consequently, these tiles will, in some cases, include several classes. Because the annotation process requires specific expertise input, the dataset contains a limited number of samples. Furthermore, the labels are imprecise as they do not include samples of the labeled border between tissue regions. This would

require multi-label samples, an even more expensive annotation process. As a result of this, the dataset is weakly labeled in both quantity and quality.

No normalization of the stain color is performed on the data, and the raw pixel intensity is used to train the models.

Stroma- and muscle-tissue are more sparsely distributed in the WSI, resulting in a smaller amount of data for these classes. Data augmentation techniques have been utilized to balance the dataset. Tiles from these two classes are extracted with 50% overlap, and further rotated and flipped during training to achieve a more balanced dataset. The size of each class is listed in Table 10.1.

Due to the low number of patients in the dataset, a traditional train/validation/test split could potentially hurt both the training and evaluation of the models. Instead, stratified 5-fold cross-validation is used. This enables the usage of all WSIs in both training and testing of the models. Stratification is performed on the patient-level to ensure that tiles from the same patient are not present in both the training and test set. A random seed is set to ensure that the folds are the same for each model, making the included samples in the training and test sets identical for all models.

Inference dataset. In addition to the CV dataset, seven WSIs were selected to be used as inference on the retrained models. The WSIs included in the inference dataset is not part of the CV dataset, and thus unseen by the models. As with the CV dataset, no normalization is performed on the WSIs in the inference dataset.

Due to the large size of the histological images, the WSIs included in the inference dataset do not have any annotations, and therefore any quantitative measurements are lacking. However, the resulting segmented images have been examined by a pathologist to be promising and confirm that the models can go from predicting smaller regions of the WSI to segment the full WSI.

10.2.2 Proposed system

An overview of the proposed system for tissue segmentation of whole slide images is presented in Figure 10.3. The system accepts input WSI of any size and outputs a corresponding segmentation image from the input. The system is tested on the seven WSIs in the inference dataset. The system consists of three main steps which will be described here. The multiscale

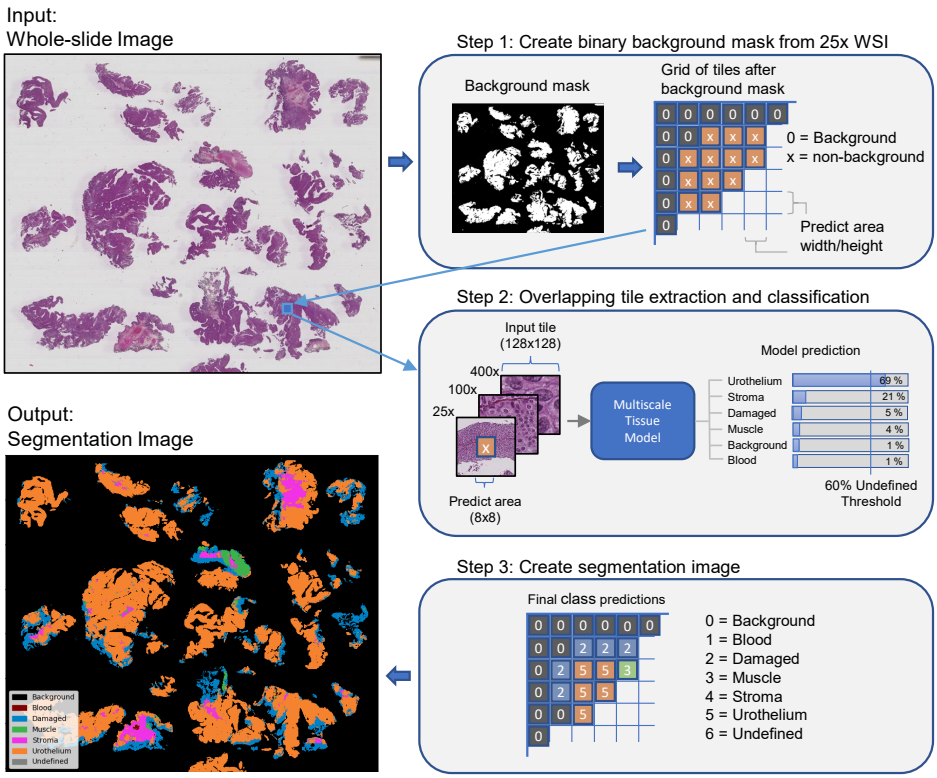


Figure 10.3: Overview of the proposed system. A background mask is created from the 25x WSI to exclude the background from further processing. Areas in the WSI selected as non-background is then extracted and fed through the multiscale model from Figure 10.4, which outputs tissue predictions. The prediction needs to exceed a set threshold to be valid. Finally, the segmentation image is generated by giving each class a separate color. The values shown in the figure are for illustration purposes only.

model in step 2 is described in more detail in the next section. Note that the blue box in step 2 in Figure 10.3 marked with 'Multiscale Tissue Model' can be exchanged with any of the models described in the model structure section below.

First, a binary background mask is produced from the 25x level of the WSI, generated by checking the pixel intensity value and splitting them into background or non-background tiles. About 60 to 80% of the WSI is covered by background, so this step reduces the number of tiles that needs to be processed by the inference model. Tiles selected as non-background are then extracted and fed to the multiscale model for further classification.

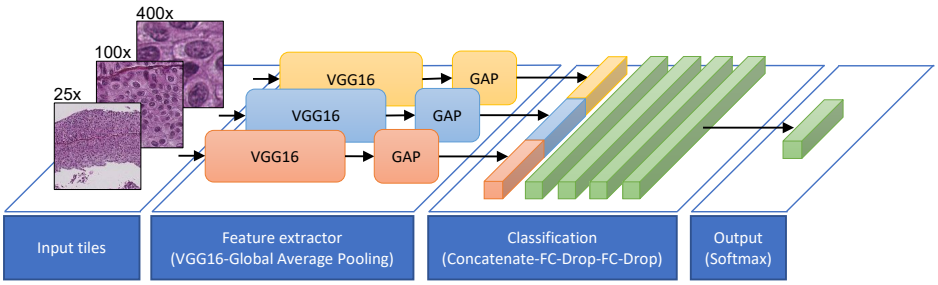


Figure 10.4: A block diagram of the TRI-CNN model proposed in the current paper. The input tiles are fed through individual pre-trained VGG16 network and global average pooling (GAP) layer to create feature vectors. The feature vectors are concatenated and fed through the classification network before entering the final output layer consisting of a softmax function. The softmax function outputs a prediction score for each of the six classes.

Depending on which model architecture is used (MONO, DI, or TRI), one, two, or three tiles are extracted from the same location but with different magnification. The extracted tile will always be 128 $\times$ 128 pixels, as this is the required input size of the inference model. However, the prediction only holds for a smaller area within the tile, typically 8 $\times$ 8 pixels, but can be set to any size. The input tiles are then overlapped, such that the inner area is located next to each other with no overlap.

Tiles are classified according to the highest prediction score. The outcome of a prediction may be equally split between multiple classes (e.g., two classes getting a score of 0.5 each, or four classes getting 0.25 each). To avoid such cases, a threshold value is set to determine if a prediction is valid. To ensure that the majority of the predicted score falls to a single class, the threshold needs to be above 0.51. Also, by setting the threshold too high may result in removing too many tiles. A threshold value of 0.6 is therefore determined as a trade-off between removing the unwanted conflicting predictions and not removing too much. Tiles with all prediction scores below the threshold are labeled as *undefined*.

Finally, each class is given a separate color, and the final segmentation image is saved. The segmentation images are ensured to only show classes with prediction scores higher than 0.6 but do not show the exact score. A method for creating heat maps has also been implemented, where no thresholding is performed, and the score for each class is visualized. A disadvantage of this is that one image must be created for each class. We earlier showed this approach in Wetteland et al. [138], but have omitted it

from this paper.

Multiscale model structure. This paper compares three architectures referred to as the MONO-, DI-, and TRI-CNN models. The three architectures have one, two, and three inputs, respectively. To differentiate the models from each other, they are named according to their main architecture, and the input scale, e.g., MONO-400x is a MONO-CNN model trained on tiles extracted at 400x magnification. Tiles in the dataset are extracted at three magnification levels, yielding three MONO models: MONO-25x, MONO-100x, and MONO-400x. These three magnification scales can further be combined in three configurations for the DI-CNN model: DI-25x-100x, DI-25x-400x, and DI-100x-400x. The TRI-CNN model has only one configuration: TRI-25x-100x-400x, and is depicted in Figure 10.4. The different MONO- and DI-CNN models can easily be derived from the same figure. E.g., to create the DI-25x-400x model, remove the 100x input and blue blocks, and to create the MONO-100x model, remove the 25x input, 400x input, red and yellow blocks.

The overall structure of each model is the same. Each input is fixed at $128 \times 128 \times 3$ pixels, which is the size of each tile. The input is fed into a pre-trained VGG16 network [113] which acts as a feature extractor, followed by a global average pooling (GAP) layer providing a feature vector representation of the input. This feature vector is then fed into a classification network consisting of two fully-connected (FC) layers, each followed by a dropout layer, and a final softmax layer with one output node for each class. The DI- and TRI-CNN models have two and three parallel VGG16 branches, respectively, resulting in multiple feature vectors. These feature vectors are concatenated before entering the classification network. The FC-layers has the same size of 4 096 neurons as the original layers in the VGG16 network. Dropout layers are added after each FC-layers to add regularization to the network due to the small dataset.

Training procedure and model selection. All models were trained using the SGD optimizer with a learning rate of 1.5e-4, batch size of 128, a dropout rate of 0.3, and a cross-entropy loss function. Early stopping was enabled, stopping the model when no increase in performance during the past 10 epochs was seen. Due to the cross-validation training scheme, no validation set was used, and the early stopping process was thus monitoring the training loss. The model is written in Python 3.5 using the Keras machine learning library [29], and Scikit-learn module [98] for evaluation.

Table 10.2: Results for all 28 models, trained using stratified 5-fold cross-validation. Each score is shown as micro-averaged F1-score aggregated across all classes, marked as 'All' in the table. F1-score only for the urothelium class is shown in the columns marked 'Uro.'. Numbers in bold refer to the highest score in their respective column.

		Multiclass				Binary-class			
		Frozen		Unfrozen		Frozen		Unfrozen	
	Model	All	Uro.	All	Uro.	All	Uro.	All	Uro.
Single scale	MONO-25x	93.4	92.9	**96.4**	96.8	96.3	92.5	98.1	96.1
	MONO-100x	94.4	96.6	94.8	97.8	98.3	96.5	99.1	98.1
	MONO-400x	87.2	89.7	86.4	86.3	94.2	88.1	93.7	87.2
Multi scale	DI-25x-100x	**96.5**	97.4	96.2	98.1	98.1	96.2	**99.3**	**98.5**
	DI-25x-400x	95.6	96.3	96.0	97.6	97.8	95.4	98.3	96.5
	DI-100x-400x	95.0	96.8	95.3	97.6	98.4	96.6	98.9	97.7
	TRI-25x-100x-400x	**96.5**	**97.6**	**96.4**	**98.3**	**98.5**	**97.0**	99.2	98.3

The models were trained in a stratified 5-fold cross-validation fashion. To produce an unbiased evaluation score, the output from each fold was summarized in a micro-average manner, as suggested by Forman and Scholz [41]. All the true positive (TP), false positive (FP), and false negative (FN) values were summarized for each class over all the folds to produce a final micro-averaged F1-score.

The VGG16 network, which is used as a base model in our architectures, is pre-trained on the ImageNet dataset [105]. It is possible to have the base model fixed during training by freezing the parameters, preventing the base model from being updated. Freezing the parameters will allow for faster training as fewer parameters need to be learned, however, as the nature of the histological images is not part of the ImageNet domain, it could affect the model's ability to fully grasp the new images. By unfreezing the weights, it may allow to better adapt to the histological domain, at the cost of longer training time. Both freezing and unfreezing the weights were tested in the experiments.

As one of the objectives is to be able to automatically extract *urothelium tissue* from the histological images, to be used in diagnostic systems in the future, it is therefore not strictly necessary to classify all six tissue classes. A possible easier problem would be to define a binary problem, classifying urothelium vs. non-urothelium tissue. Each model was therefore also tested with this binary-class approach to see if it improved classification results for urothelium tissue. By simply combining the remaining five classes into one non-urothelium class, the dataset becomes heavily unbalanced towards the non-urothelium class. To counteract against this, augmentation using rotation and flipping was applied to balance out the dataset. By augmenting all the tiles from the muscle, stroma, and urothelium class 4x during training, the dataset became evenly distributed between the two classes urothelium and non-urothelium.

After evaluating the model using stratified cross-validation, a new and final inference model was trained by utilizing all available data as training data. The average number of epochs used during cross-validation was used when training the inference model. This inference model was then used to predict new WSIs from the inference dataset.

10.3 Results

This section will present the results for the different models. A total of 28 models were trained using stratified 5-fold cross-validation, including

single- and *multiscale*, and *binary-* and *multiclass* models. Each model was trained using weakly labeled data, with both frozen and unfrozen weights in the VGG16 network.

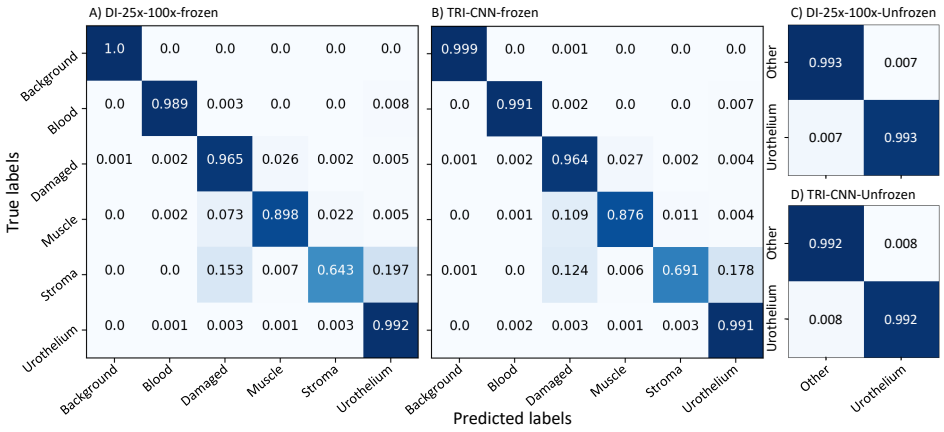


Figure 10.5: Normalized confusion matrices for the best multiscale models. Aggregated results across all five folds in the cross-validation test. A) Best multiclass DI-CNN, B) Best multiclass TRI-CNN, C) Best binary-class DI-CNN, and D) Best binary-class TRI-CNN.

Table 10.2 shows the cross-validation results for all the models. Aggregated micro-average F1-score across all classes are included, as well as the F1-score for only the urothelium class to better compare multiclass vs. binary-class models. Figure 10.5 displays the confusion matrices for the best multiclass models. The matrices are normalized to allow for more easy comparison. For the number of samples in each class, refer to Table 10.1.

Some of the best models have been retrained on the entire CV dataset and used to segment the seven WSIs included in the inference dataset. The resulting segmented images have then been inspected by an expert pathologist and are considered to be very promising. Figure 10.6 shows four WSIs and their corresponding tissue segmented images generated by the best multiclass model. Figure 10.7 shows a comparison between segmentation images generated by the best binary-class model and the best multiclass model. A DICE-score is calculated to measure the similarity between the predicted urothelium tissue between these two models, with an average DICE-score of 0.87 for the three WSIs. Figure 10.8 shows a close-up region taken from the top-right corner of the first WSI in Figure 10.6. This region is then segmented with all the best MONO-, DI-, and TRI-models for comparison.

10.4 Discussion

The results in Table 10.2 are shown as micro-averaged F1-score across all classes, as well as for the urothelium class. The results are overall good for all models, and a discussion of each case follows below. Afterward, the confusion matrices and the segmented images will be discussed, and finally, different usage scenarios of the system will be considered as well as some limitations of the study.

Binary-class vs. multiclass. As expected, the binary-class models achieve a higher average F1-score than the multiclass models, with all 14 of the binary models getting a higher score than their multiclass counterparts. This is expected because five of the classes are now grouped, and misclassification within these classes is canceled out. The best multiclass model is the frozen TRI-25x-100x-400x with an F1-score of 96.5% across six classes, whereas the best binary model is the DI-25x-100x with unfrozen weights, which got an F1-score of 99.3% across its two classes.

By looking at the F1-score for the urothelium class alone, the multiclass models are now superior, with 9 of the 14 models being ahead of their binary-class counterparts. The few binary-models which have a higher score, are only marginally so, with the largest difference being the unfrozen MONO-400x, where the binary version is 0.9% better than the multiclass version. It is clear that by simplifying the problem into a two-class problem, did not help with getting better urothelium extraction. The highest urothelium score is achieved by the TRI model, where both the unfrozen multiclass and unfrozen binary-class version each got an equal F1-score of 98.3% for the urothelium class.

Frozen vs. unfrozen. The three architectures MONO, DI and TRI, have 19M, 21M, and 23M trainable parameters, respectively, when the VGG16 weights are frozen. By unfreezing the weights, the same models get 34M, 50M, and 67M trainable parameters. When comparing results for these models, there is on average an increase of +0.6% by unfreezing the weights. Of the 14 unfrozen models, 10 get a higher score than the corresponding frozen models. The largest increase is seen in the binary MONO-25x model, which goes from an F1-score of 96.3% to 98.1% by unfreezing the weights.

The increase in the number of trainable parameters also affects the training time of the models. The average time per epoch for all the frozen models was 9 minutes, while the unfrozen models needed on average 10 minutes to compute one epoch. This is an increase of 11% processing time per epoch. However, the frozen models needed on average 162 epochs to reach the early stopping criteria, whereas the unfrozen models only needed 58 epochs. Thus, the models with unfrozen weights needed about 60% less processing time during training.

Single-scale vs. multiscale. When comparing the single-scale MONO-models with the multiscale DI- and TRI-models, the multiscale models achieve better results across all columns in Table 10.2, with the exception for the unfrozen MONO-25x model which matches the performance of the TRI-scale model. If we limit ourselves to the multiclass models, the best models for the three architectures are the unfrozen MONO-25x with 96.4%, frozen DI-25x-100x with 96.5%, and frozen TRI-25x-100x-400x which got an F1-score of 96.5%. The story is similar for the binary models, with unfrozen MONO-100x being the best with 99.1%, unfrozen DI-25x-100x with 99.3%, and unfrozen TRI-25x-100x-400x with 99.2%.

By looking at the single-scale models alone, it is clear that the two lower scales (25x, 100x) are performing better than the 400x scale, and that having a greater field of view is preferable. The multiscale models, consisting of two and three VGG16 networks, have a more complex structure involving more parameters than the MONO models. In addition, they have access to a greater field of view in all its models. These two features seem to help the performance of these models.

Naturally, the MONO models take the least amount of training time, with an average of 4:40 minutes per epoch. The DI-models take 136% longer with an average of 11:01 minutes, and finally, the TRI-models take the most time with 19:38 minutes on average per epoch. That is 321% and 78% longer than MONO and DI, respectively. The average number of epochs before reaching the early stopping criterion for the three architectures was 147, 88, and 64 epochs for the MONO-, DI-, and TRI-models, respectively.

Confusion matrices. Figure 10.5 shows the resulting normalized confusion matrices for the best multiscale models for both multiclass and binary-class models.

In the two multiclass matrices (A) and (B), the models did an excellent job at classifying background, blood, and urothelium correctly, and a great job with the damaged class as well. Both models struggled mostly with the muscle and stroma classes. These are the classes with the fewest number of labeled samples in the dataset. As a result of this, the models may have achieved a weaker generalization for these classes, and thus misclassified them more often. Most notable misclassifications are related to muscle and stroma being misclassified as damaged tissue, and also stroma being misclassified as urothelium.

The two binary-class models in Figure 10.5 (C) and (D) got an equally good performance. Five of the classes are now combined into one class named *other* in the figure and thereby removing most of the misclassifications from the multiclass cases. However, this did not significantly increase the performance of model (C) and (D). Model (D) got the same normalized score as (A), and model (C) is only marginally better.

Inference dataset results. The seven WSIs included in the inference dataset were processed with overlapping tiles according to Figure 10.3, where only the inner 16 $\times$ 16 pixel of the tile was classified. The average processing time was 7 hours 18 minutes, including all three steps in Figure 10.3. On average, only 0.9% of the WSIs were categorized as undefined. Four of the WSIs are presented in Figure 10.6 and three in Figure 10.7.

Segmentation image results. The best multiclass model, according to Table 10.2, is split between two models. The frozen DI-25x-100x and frozen TRI-25x-100x-400x both have a similar F1-score of 96.5%, but the latter model has a higher urothelium F1-score and is thus regarded as the best multiclass model. The model was retrained and used to process four new WSIs, not present in the training data, to demonstrate its usage. Figure 10.6 shows the original WSI with the corresponding segmentation images. The segmented images are intuitive, easy to understand, and allow even untrained personnel to both identify and locate the difficult to find regions, e.g., like muscle tissue.

Fully multiclass-annotated WSI in our dataset is not available. The resulting segmentation images for the WSI have, however, been manually inspected by an expert uropathologist and are considered to be very promising, especially considering that the WSIs were only weakly annotated. Large homogeneous areas with a certain tissue type are clearly recognized. Most models are really challenged by smaller, more heterogeneous areas.

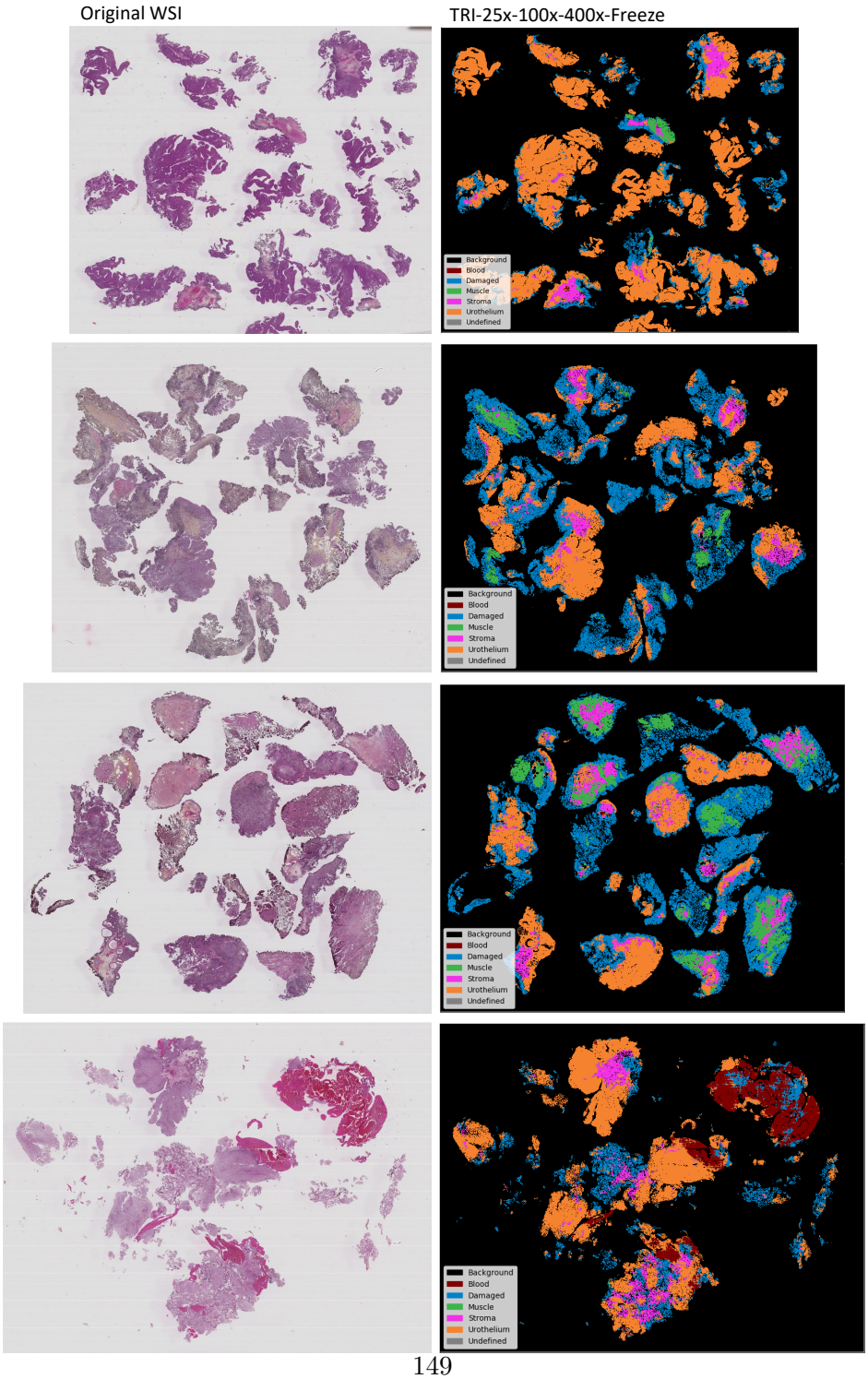


Figure 10.6: The best multiclass model was retrained and used to generate segmentation images from four WSI not present in the training data.

Binary-class vs. multiclass segmentation images. The best Multiclass and binary-class models were retrained and used to create the segmentation images seen in Figure 10.7. The multiclass segmentation image may be of more interest to a pathologist, as it outlines regions of all six classes, whereas the binary-class segmentation image only outlines the urothelium class. However, both the multiclass and binary-class models have about the same F1-score for the urothelium class, and the additional information in the multiclass segmentation images favor the former model in a final system.

After comparing the urothelium regions in the two segmented images for each WSI, they are very similar. The DICE-score is calculated to measure the similarity between the regions, and the three cases have an average DICE-score of 0.87, which confirms that the two model's prediction for urothelium is quite similar. However, there is no truth annotation, so the DICE-score does not reveal if one of the models is better than the other.

Close-up segmentation regions. Even though the system is trained on weakly labeled data, consisting of single-class samples, using tile-based classification and not a per-pixel classification, it is still interesting to see how the system performs on a detailed level. This also allows us to compare the different models. Figure 10.8 shows a close-up region taken from the top-right corner from the first WSI in Figure 10.6, processed using an 8×8 pixel predict area.

All models do a decent job of outlining the major regions in the image. The different models process the image on different scales, and so the prediction tile covers a larger area for the smaller scales. The effect of this is visible at the three MONO models, where the level of detail goes up with each scale. The MONO-100x and MONO-400x models, with its smaller field of view, are able to detect some of the small regions containing blood in the middle of the image. The MONO-25x, however, is not able to identify this. The DI-25x-100x model, which has access to both the mid and broad field of view, barely identifies a small part of the blood, whereas the TRI scale model does not identify it at all.

10.4.1 Usage scenarios

As seen from both Table 10.2 and the segmented images in Figure 10.6, the model is fully capable of distinguishing between the different tissue types.

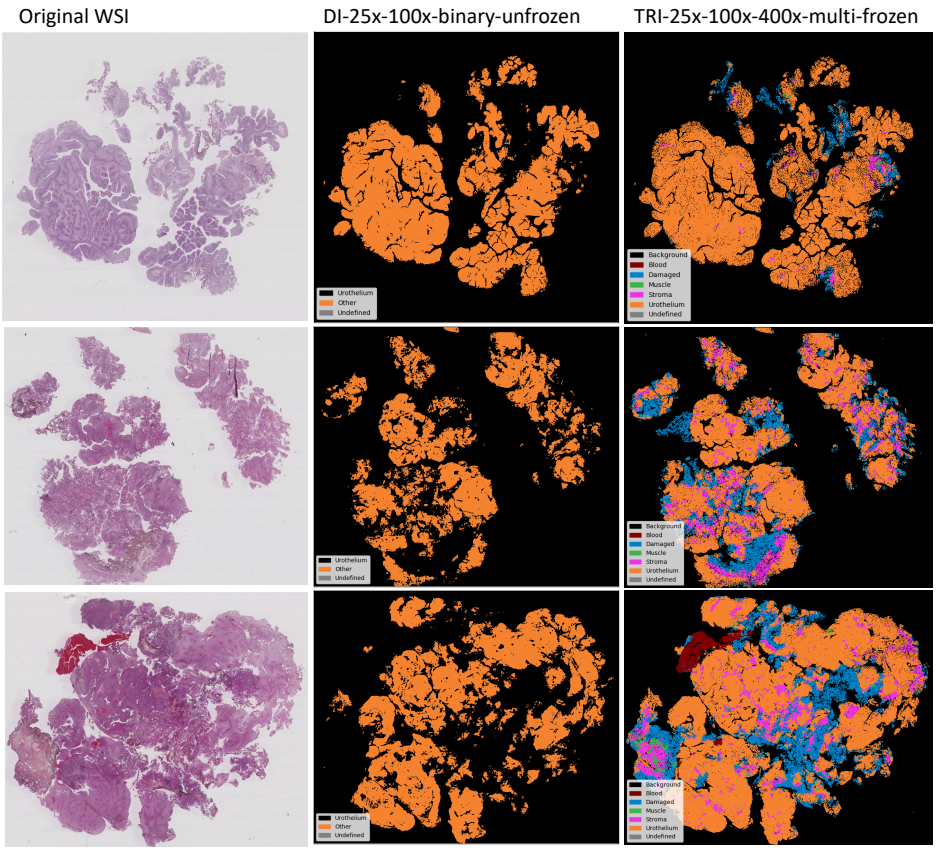


Figure 10.7: The best binary-class model vs. the best multiclass model. A DICE-score is calculated to measure the similarity between the predicted urothelium tissue between the two models. DICE-score from top to bottom are 0.92, 0.85 and 0.85.

The presented system has several possible usage scenarios, which will be discussed here.

The segmented images in Figure 10.6 can be used as a digital tool for pathologists to help them become more efficient in their work. It can be used to guide them to the diagnostic relevant areas of the WSI, such as urothelium, muscle, and stroma tissue. It can also be used to find edges of the urothelium tissue without damage more easily. During an examination, a pathologist needs to verify if muscle tissue is present or not in the current WSI. With the segmented images, this can be verified within a short amount of time.

Another use case for the system is as a preprocessing step for an automatic

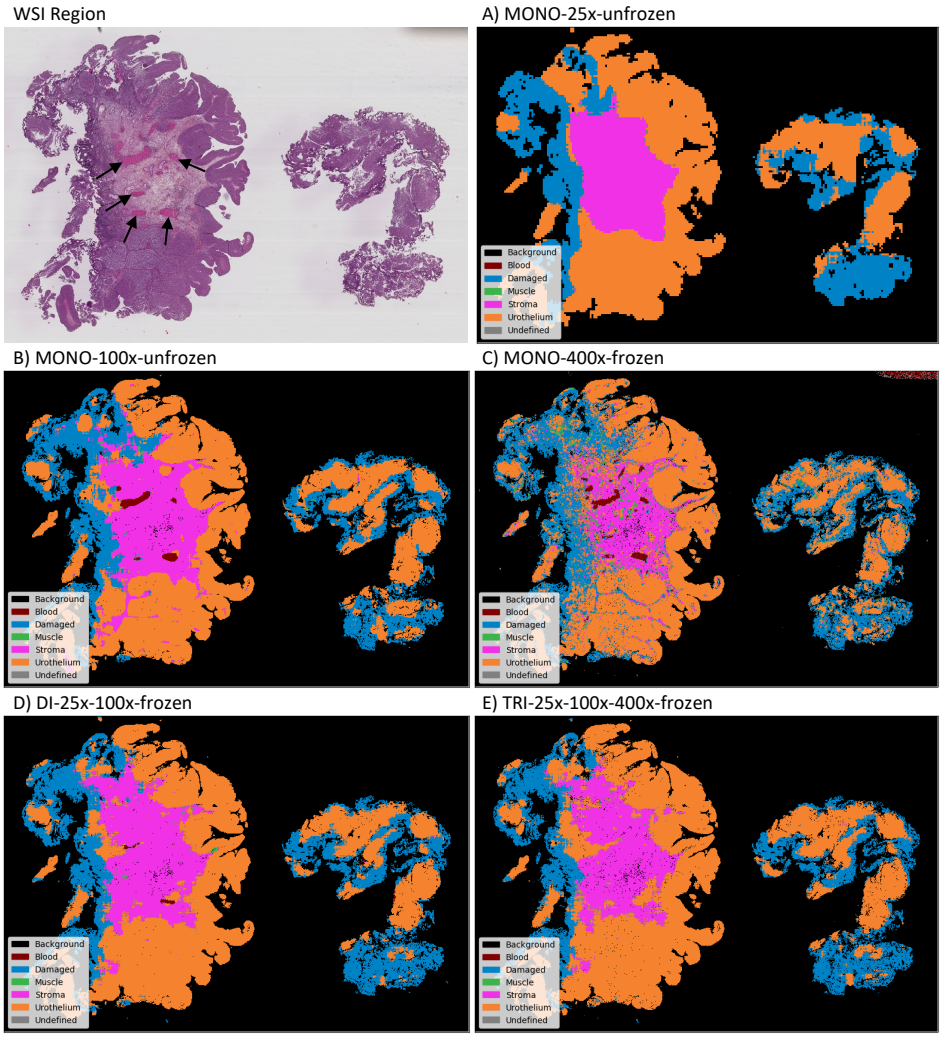


Figure 10.8: Segmentation of close-up region taken from the top-right corner from the first WSI in Figure 10.6. A) Best MONO-25x, B) Best MONO-100x, C) Best MONO-400x, D) Best DI-CNN model, E) Best TRI-CNN model. Arrows in the WSI region points to small areas of blood that the models struggle to identify.

diagnostic system. For instance, each patient has follow up records about whether the patient experienced recurrence and progression. By training a diagnostic model on the entire WSI, the dataset quickly becomes too large if many patients are included. Also, by randomly selecting a subset of tiles within each WSI, the dataset will include a large portion of damaged tissue and blood, which will add noise to the diagnostic model. By using the multiscale tissue model presented in this paper as a preprocessing step, areas of clean, undamaged urothelium and other diagnostic relevant types can easily be extracted and used as training data.

10.4.2 Limitations

One limitation of the current study is that the dataset is relatively limited in size. A small training dataset may lead to overfitting of the model, resulting in poor performance, and a small test set may cause an optimistic estimate of the performance. Several measures have been taken to reduce these negative effects. Pre-trained models, dropout, and early stopping was used to reduce overfitting, and cross-validation was used to get a realistic estimate of each model's performance.

As mentioned in the data material section, the labels are accurate in the highest resolution (400x) but are imprecise on the lower scales (25x, 100x), meaning the ground-truth is based on weak annotations of the dataset, which may impact the accuracy. The experimental results show that having access to a greater field of view outweighs the potential negative effects of imprecise labels.

It is difficult to compare the presented models against other approaches or to perform a test on an independent dataset. To the best of the authors' knowledge, no other open dataset exists with annotations of the same six classes. As mentioned in the related work section, some research and models exist for segmentation of histological images. However, these are based on other cancer types or trained on other classes than the six classes used in this paper.

10.5 Conclusion

This paper investigates the effect of using multiple scales during tissue classification from WSI of urothelial carcinoma into six classes. The classification is performed on smaller tiles and can be useful for a coarse

segmentation, or ROI-extraction, of WSI. Three main architectures are presented: MONO-, DI-, and TRI-CNN model, and a total of 28 different models were trained using weakly labeled data and evaluated in a stratified 5-fold cross-validation scheme.

The multiscale models achieved a better result than the MONO-CNN models. There was not a substantial increase in urothelium classification by using the binary-class models, neither by cross-validation or by inspection of the segmented images. The best multiclass model was used to generate intuitive and easy to understand segmented images from unseen WSIs, and after inspection by a pathologist is considered to be very promising.

The segmented regions shown in Figure 10.8 demonstrates the importance of including the highest magnification scale (400x) during tile-wise classification. The models which do not include this scale are not able to identify the smaller details within the WSI.

As the three MONO models pick up different levels of details, we will in the future experiment on employing them in a multiscale ensemble model by combining their outputs, instead of combining the different scales within the models, as the DI- and TRI-CNN models do. We also plan to use the model for automatic ROI-extraction of relevant tissue in the WSI to create training datasets for a diagnostic and prognostic classification model. By only extracting the diagnostic relevant areas of the WSIs, a dataset of much higher quality can be collected.

Authors' note

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. As this is a retrospective study, Ethical approval was given without written consent from the patients. All insights in a patient's journal are monitored electronically, and all except the treating physician were required to state the reason why they needed to read that patient's journal. This log is always open for the patient to view. All patients were checked if any had registered themselves in the register for research reservation from the National Institute of Health (Registry of Withdrawal from Biological Research Consent, Norway).

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research is partly funded by the University of Stavanger, Faculty of Science and Technology, Strategic PhD scholarship in health technology.

Paper 3: Semi-Supervised Tissue Segmentation of Histological Images

Semi-Supervised Tissue Segmentation of Histological Images

Ove Nicolai Dalheim[1,4], Rune Wetteland[1,4], Vebjørn Kvikstad[2,3], Emiel A. M. Janssen[2,3], Kjersti Engan[1]

[1] Department of Electrical Engineering and Computer Science, University of Stavanger, Norway

[2] Department of Chemistry, Bioscience and Environmental Engineering, University of Stavanger, Norway

[3] Department of Pathology, Stavanger University Hospital, Norway

[4] These authors contributed equally to the work

Published in the Proceedings of the 10th Colour and Visual Computing Symposium (CVCS), 2020.

http://ceur-ws.org/Vol-2688/

Abstract:
Supervised learning of convolutional neural networks (CNN) used for image classification and segmentation has produced state-of-the-art results, including in many medical image applications. In the medical field, making ground truth labels would typically require an expert opinion, and a common problem is the lack of labeled data. Consequently, the models might not be general enough. Digitized histological microscopy images of tissue biopsies are very large, and detailed truth markings for tissue-type segmentation are scarce or non-existing. However, in many cases, large amounts of unlabeled data that could be exploited are readily accessible. Methods for semi-supervised learning exists, but are hardly explored in the context of computational pathology. This paper deals with semi-supervised learning on the application of tissue-type classification in histological whole-slide images of urinary bladder cancer. Two semi-supervised approaches utilizing the unlabeled data in combination with a small set of labeled data is presented. A multiscale, tile-based segmentation technique is used to classify tissue into six different classes by the use of three individual CNNs. Each CNN is presented tissue at different magnification levels in order to detect different feature types, later fused in a fully-connected neural network. The two self-training approaches are: using probabilities and using a clustering technique. The clustering method performed best and increased the overall accuracy of the tissue tile classification model from 94.6% to 96% compared to using supervised learning with labeled data. In addition, the clustering method generated visually better segmentation images.

11.1 Introduction

In Norway, 1 748 patients were diagnosed, and 319 people died from bladder cancer in 2018. The majority of these, at 73%, were male, while the remaining 27% were female [64]. Worldwide in 2018, 199 922 people of both sexes died of bladder cancer [123], and 549 393 new patients were diagnosed, placing bladder cancer as the 10th most common cancer type in the world. Since 2001, bladder cancer (including the urinary tract) has been the fourth most common cancer diagnosis for men in Norway [20, 21, 22, 23]. In addition, bladder cancer is known as one of the most recurring cancer types, with the probability of recurrence for high-risk patients after one year at 61% [3].

An important step in determining the cancer stage and correct treatment plan for bladder cancer patients is to examine the tissue samples that are extracted during transurethral resection. The tissue samples contain large amounts of information from individual cell characteristics, to specific cell quantities in large tissue clusters. Scanning and digitalization of the histological stains produce whole slide images (WSI), uncovering the field of computational pathology. A significant increase is seen in the number of tissue samples sent to pathologist labs, affecting the waiting time for patients [116]. The increase in amount of specimens is unfortunately not seen in the number of pathologists. Another aspect is that since the WSI is studied manually, pathologists staging and grading of bladder cancer may differ in relation to the same tissue as pathologists have a different set of subjective expectations and experiences. With computational pathology, computerized tools can aid the pathologist in diagnostic predictions, localization of interesting regions, and segmentation, to name a few applications.

During the last decade, convolutional neural networks (CNN) have proven very useful in image processing and image classification tasks [42, 147]. CNNs are gaining popularity also in medical image processing and in computational pathology. The most common way to train neural networks (NN) is by supervised learning (SL) and backpropagation. This requires a large training set where all samples have associated relevant ground truth labels. Labeled data within medicine is often limited, and producing it is a time-consuming process that requires annotations made by experts. A way around the lack of labels is clustering or unsupervised learning. One method is the use of autoencoders, where a compression-decompression setup is used, making the network try to reconstruct the original input [18]. The learned features are found at the most compressed state, and might

ultimately be connected to a classification network. The drawback here is that they rarely perform as well as models trained with a supervised method.

CNNs are referred to as shift-invariant, meaning that a particular feature can be detected wherever it may be located in the image. Intuitively, the initial layers of a CNN can be viewed as feature extraction, while the last layers can be viewed as the most task-specific object detection or classification layers. There are many parameters to go about when setting up a new CNN, and normally large quantities of labeled data are needed to do so. Therefore, the first layers can be inherited from a pre-trained network, and the last layers are trained from scratch, a process known as transfer learning [94].

A consolidation of the above methods is semi-supervised learning (SSL), where both labeled and unlabeled data is used to train a network. This can be beneficial in cases where there are small amounts of labeled data, but large quantities of unlabeled data. Different semi-supervised methods exist, like graph-based learning methods that often implement clustering algorithms to locate and distinguish inputs in feature space [143]. One other semi-supervised method called self-training aims to first train a NN on labeled data in a supervised manner. Thereafter, predictions are found for new unlabeled data using the first model, and finally, a new model can be trained on both the ground truth labels from annotations and the weak labels from the predictions [129].

In very recent years, we find some works on semi-supervised learning within computational pathology. In Dercksen et al. [61], a method based on autoencoders and k-means clustering of features is presented. A combination of contrastive predictive coding and multiple instance learning on breast cancer data is presented in Lu et al. [77]. In Peikari et al. [99], a cluster-then-label approach is taken using SVM classifiers. Our group presented a method for multiclass tissue classification of urothelial carcinoma in [138, 139, 140]. Encouraged by the results, but challenged by the lack of labeled data to generalize the model further and utilize larger amounts of unlabeled data, we propose to combine the TRI-CNN transfer-learning based architecture with semi-supervised learning.

This paper presents two methods within self-training applied to tissue segmentation of WSIs of urothelial carcinoma. The first method is a probability-based method based on predicted probabilities from an initial model. The second method is a cluster-based self-training method based on

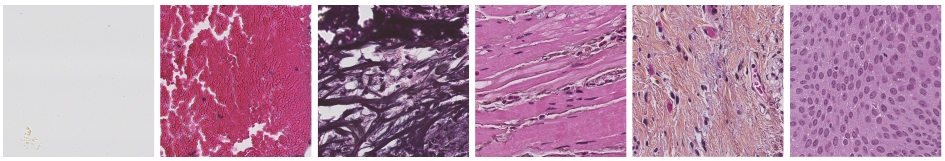


Figure 11.1: Tissue representing the different classes used in the AI models. From left to right: Background, Blood, Damaged, Muscle, Stroma, Urothelium.

both predicted probability from the initial model and local neighborhood in the predictions.

11.2 Material and methods

11.2.1 Data material

The material used in this paper consists of tissue samples from tumors of patients with bladder cancer in the form of urothelial carcinoma. The tumor is removed from the patient through Transurethral Resection of Bladder Tumor (TURBT) by the use of a resectoscope. The resectoscope holds a heated wire loop for removing the tumors, and the resulting tissue will often bear marks with burnt or torn tissue. After the tumor is removed, it is fixed in formalin before being embedded into paraffin. When the paraffin is solidified, it has a similar consistency to tissue and can more easily be sliced into 4 µm thick slides with a microtome. Variation in slice thickness can occur, in turn sourcing problems like color variation and tissue folds in the resulting image, opposing and extra challenge to the classifier. The slices are then stained with Hematoxylin Eosin Saffron [36] and further scanned with the digital slide scanner system, Leica SCN400, to produce the WSI. This, as well as previous work done on the same dataset, leads to the six classes which can be seen in Fig. 11.1.

The manually marked ground truth dataset, D_{gt}, consists of 37 patients, from which 125 020 tiles have been extracted. The labels originate from annotations made at 400x magnification by a pathologist at Stavanger University Hospital, (VK), illustrated in Fig. 11.2. It is a private dataset, however, reasonable requests may be made to the corresponding author. The three extracted tiles have the same size of 128 $\times$ 128 pixels, but are extracted at different magnification levels. The lower magnification tiles (25x, 100x) have a larger field-of-view than the high magnification

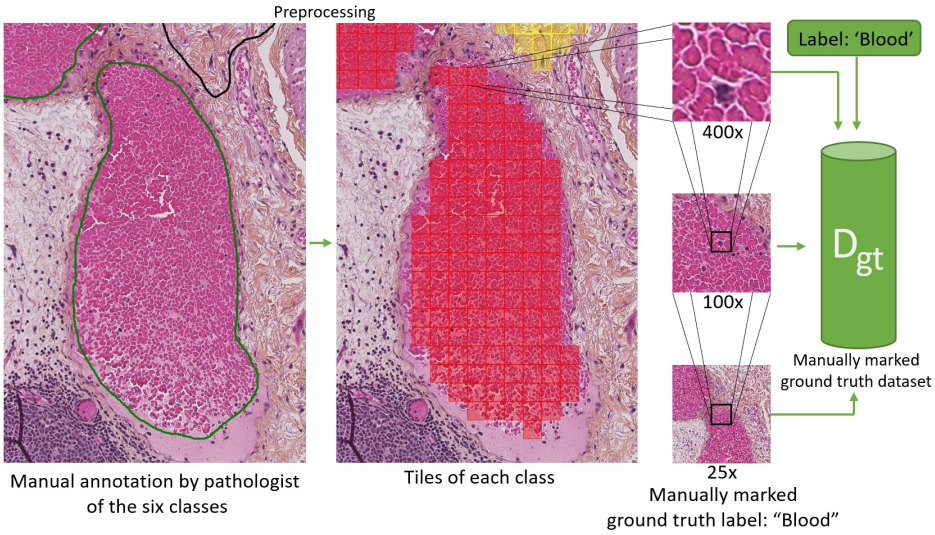


Figure 11.2: Origin of manually marked ground truth dataset, D_{gt}.

tile (400x), allowing the multiscale model to capture both details and context of the input images. The coordinates are then saved with the three magnification levels, accompanied by its corresponding ground truth label. The dataset was divided into $D_{gt}\{train\}$ consisting of 103 650 tiles from 29 patients, and $D_{gt}\{test\}$ consisting of 21 370 tiles from 8 patients.

46 new patients from the unlabeled dataset were chosen to extract tiles from, with the two self-training methods. For the probability-based method, a total of 121 239 tiles were extracted from all 46 patients and formed the probability-weak dataset, D_{pw}. For the cluster-based method, a total of 221 612 tiles were collected from 44 patients and formed the cluster-weak dataset, D_{cw}. An overview is presented in Table 11.1.

11.2.2 Methods

This section presents the original model, which originates from the framework developed by Wetteland et al. [140]. Afterwards, the methods behind the two self-training approaches within semi-supervised learning are explained.

Table 11.1: Overview of labels used during training of the different models.
Ba = Background tiles, Bl = Blood tiles, Da = Damaged tissue tiles,
Mu = Muscle tissue tiles, St = Stroma tissue tiles, Ur = Urothelium tiles.

Label type	**Dataset**	**Ba**	**Bl**	**Da**	**Mu**	**St**	**Ur**	**Total**
Ground truth	$D_{gt}\{train\}$	21 423	16 949	28 452	8 061	3 614	25 151	103 650
Ground truth	$D_{gt}\{test\}$	5 589	2 883	5 155	1 905	1 261	4 577	21 370
Probability-weak	D_{pw}	20 300	20 036	20 176	20 416	20 229	20 082	121 239
Cluster-weak	D_{cw}	21 281	42 630	24 817	48 359	52 794	31 731	221 612

Initial supervised approach

The original model arises from a traditional supervised learning method, using the ground truth labels presented in Table 11.1. The dataset, D_{gt}, is split between a train/test ratio of approximately 83/17, taking into account that the same patient does not exist in both sets regardless of class. The individual per-class train/test split varies from a 86/14 ratio for blood to a 74/26 ratio for stroma. All models trained using a SL approach are referred to as TRI-SL.

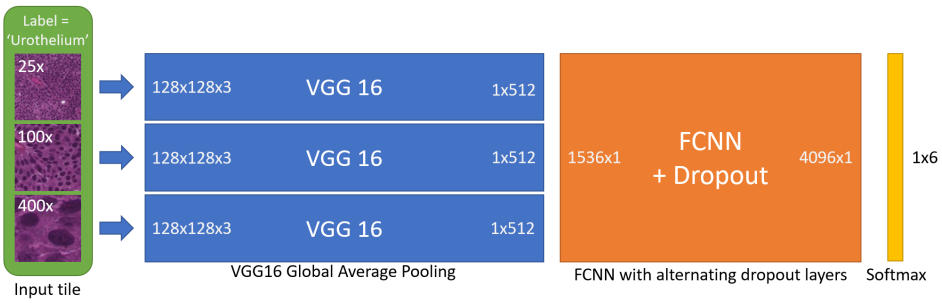


Figure 11.3: Illustration of the multiscale TRI-architecture used in all models.

As illustrated in Fig. 11.3, the architecture of the TRI-CNN model utilizes transfer learning by implementing three VGG16 models [113] in parallel that operate individually. The VGG16 network converts a 128 $\times$ 128 $\times$ 3 input RGB image into a feature vector with dimension 1 $\times$ 512. This is done by a sequence of five CNN blocks that each consist of two or three CNN layers followed by a rectified linear unit (ReLU) layer and finally an average pooling layer. The three 1 $\times$ 512 outputs from the VGG16 models are then merged into a single 1 $\times$ 1 536 layer followed by a fully-connected neural network (FCNN). The FCNN consists of two layers with 4 096 neurons each, with one dropout layer between them. Thereafter, another dropout layer before the final output layer classifies the tissue with a Softmax activation function. Each VGG16 network is fed the input tiles at the three different magnifications 25x, 100x and 400x, to allow for different features to be detected at each level. The multiscale model is therefore abbreviated with the name TRI-CNN, which originates from the nomenclature in Wetteland et al. [140].

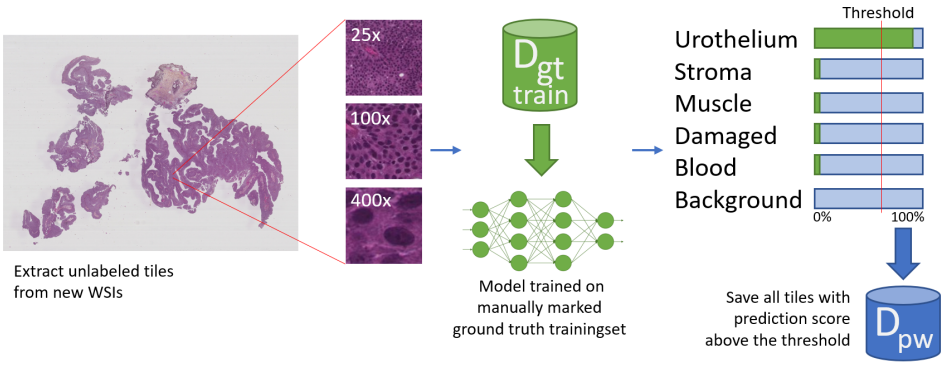


Figure 11.4: Origin of probability-weak dataset, D_{pw}.

Probability-based self-training

The probability-based self-training method is the most straight forward approach within self-training. Each of the 46 images is split up into tiles of size 128 × 128 pixels, and each tile is classified by the original model, TRI-SL-AF, which is trained on the ground truth labels. Every tile that is classified with a minimum probability threshold of 60% is saved, while tiles classified with lower probability are discarded. The 60% threshold is a trade-off between acquiring enough tiles while having a large enough probability. As illustrated in Fig. 11.4, the saved tiles are then selected based on several criteria given in Table 11.2. All models trained using the probability-based self-training method are referred to as TRI-P-SSL.

Table 11.2: Tile criteria for probability-weak dataset D_{pw}.
Ba = Background tiles, Bl = Blood tiles, Da = Damaged tissue tiles,
Mu = Muscle tissue tiles, St = Stroma tissue tiles, Ur = Urothelium tiles.

Criteria	Ba	Bl	Da	Mu	St	Ur
Min. tile probability	95%	80%	95%	95%	95%	95%
Max. tiles per WSI	1 900	8 000	710	5 000	5 000	480
Min. tiles per WSI	707	53	277	707	5 688	32 916
Max. tiles tot.	20 500	20 500	20 500	20 500	20 500	20 500

The method used to select tiles from the 46 patients is designed to select the tiles only based on its probability score across all WSIs. First, a scan runs through all the patients and counts the number of tiles per patient. Patients with an insufficient number of tiles according to the minimum

number of tiles per WSI are discarded, and tiles are collected from the remaining patients. For each patient, tiles with the highest probability are collected first, until the maximum number of tiles per WSI has been collected, or no more sufficient tiles remain. All tiles from all WSIs are then appended to an array and sorted based on probability. The tiles with the highest probability are then selected from this array according to the maximum total number of tiles. This is done for each class and later saved to the probability-weak dataset D_{pw}, see Table 11.1.

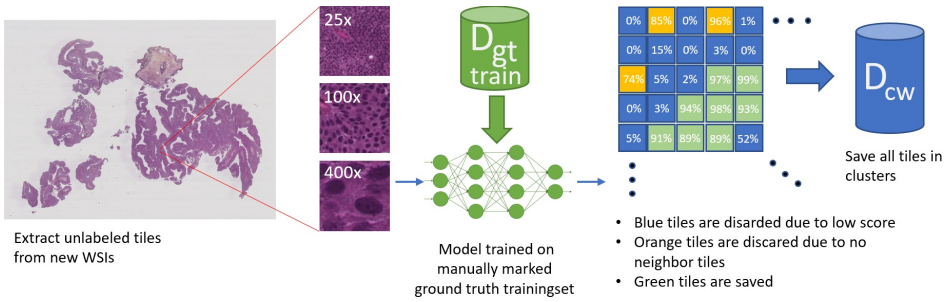


Figure 11.5: Origin of cluster-weak dataset, D_{cw}.

Cluster-based self-training

Similar to the probability-based method, the cluster-based method uses model TRI-SL-AF to classify the WSIs. The tiles are classified with a minimum of 60% probability, and tiles with a lower probability are discarded. The classified tiles are then selected based on several criteria listed in Table 11.3. A visual representation of this is illustrated in Fig. 11.5. All models trained using the cluster-based self-training method are referred to as TRI-C-SSL.

An algorithm searches through the tiles and groups them into clusters. If, at any point in the search, the maximum number of tiles per cluster is not reached, the difference is appended to the limit of the next cluster in line. The average cluster probability is calculated per cluster, and the clusters are sorted after the highest probability. Each cluster originating in the WSI is then sorted into an array, and the program selects the clusters based on the highest probability according to the maximum number of clusters. The labels are then saved to the cluster-weak dataset D_{cw}, see Table 11.1.

Table 11.3: Tile criteria for cluster-weak dataset D_{cw}.
Ba = Background tiles, Bl = Blood tiles, Da = Damaged tissue tiles,
Mu = Muscle tissue tiles, St = Stroma tissue tiles, Ur = Urothelium tiles.

Criteria	**Ba**	**Bl**	**Da**	**Mu**	**St**	**Ur**
Min. tiles per WSI	50	20	50	20	50	50
Max. tiles per WSI	20 000	20 000	798	4 815	1 440	1 235
Max. clusters per WSI	100	100	100	100	100	100
Min. cluster size	50	20	50	20	50	50
Max. tiles per cluster	20 500	20 500	20 500	20 500	20 500	20 500
Min. avg. cluster prob.	60%	60%	60%	60%	60%	60%

11.3 Experimental setup

Six multiscale models are presented in this paper, and the following letters are used to describe them: SL is short for supervised learning, and SSL for semi-supervised learning. P indicates that the models are trained through the probability-based self-training method, and C implies that the cluster-based self-training method is used. A refers to that augmentation by rotation of tiles is involved. F and U refer to the weights in the VGG16 models being frozen or unfrozen during training, respectively. An overview is given in Table 11.4.

Models TRI-SL and TRI-SL-AU were trained through supervised learning on dataset $D_{gt}\{train\}$ and tested on $D_{gt}\{test\}$, see Table 11.1. The models based on the probability-based self-training method, TRI-P-SSL and TRI-P-SSL-AU, were trained on the labels in both $D_{gt}\{train\}$ and D_{pw}. TRI-C-SSL and TRI-C-SSL-AU were trained with the cluster-based self-training method on labels from both datasets $D_{gt}\{train\}$ and D_{cw}. The models TRI-SL-F, TRI-P-SSL-F, and TRI-C-SSL-F were trained with VGG16 frozen, meaning only the FCNN and output layer was trained. For models TRI-SL-AU, TRI-P-SSL-AU, and TRI-C-SSL-AU, the VGG16 model was unfrozen during training, and weight in the whole network was adjusted.

For the original model, TRI-SL-F, stroma and muscle tissue tiles were augmented by rotation to produce two times as many tiles in an effort to equalize the dataset with respect to tiles per class. For models TRI-P-SSL-F and TRI-C-SSL-F, no augmentation was used. Models TRI-SL-AU, TRI-P-SSL-AU, and TRI-C-SSL-AU were all trained with 3x augmentation of tiles in all classes except background, as the background is filtered out

Table 11.4: Overview of the six models. st = stroma tiles, mu = muscle tiles

Model	Magnification	Method	Augm.	VGG16
TRI-SL-AF	25x, 100x, 400x	Supervised learning	2x st, mu	Frozen
TRI-P-SSL-F	25x, 100x, 400x	Probability-based SSL	No	Frozen
TRI-C-SSL-F	25x, 100x, 400x	Cluster-based SSL	No	Frozen
TRI-SL-AU	25x, 100x, 400x	Supervised learning	3x	Unfrozen
TRI-P-SSL-AU	25x, 100x, 400x	Probability-based SSL	3x	Unfrozen
TRI-C-SSL-AU	25x, 100x, 400x	Cluster-based SSL	3x	Unfrozen

such that only the foreground is processed by the models. This is done to save processing time, however, background tiles containing debris were not filtered out, and needs to be processed.

During training of all six models the learning rate was set to 1.5e-4 at a batch-size of 128. The stochastic gradient descent (SGD) backpropagation algorithm was used as optimizer, and the dropout rate was set to 20%. An early-stopping criterion was set to end training when the change in validation loss was smaller than 1e-6 for six consecutive epochs. No weighting of the different labels in the datasets was used during training. All methods were implemented in Python 3.5, with TensorFlow 1.13 [1] and Keras 2.3 [29]. Scikit-learn [98] was used for evaluation, and PyVips [84] was used to process the images.

11.4 Results

All six multiscale models were tested on dataset $D_{gt}\{test\}$, yielding the results in Table 11.5. To further investigate the individual model performance with regards to segmentation, a new WSI was segmented by all six models by tile-wise classifying all foreground regions without overlap. The WSI has been annotated by a pathologist and has not been used during training before. This WSI is referred to as WSI_segment_test, and the predictions of the WSI is compared to the ground truth annotations in it. Fig. 11.6 shows the close-up 400x image of an area in WSI_segment_test, where the whole foreground is labeled as blood, with the corresponding prediction by all six models. A visual comparison of an area in WSI_segment_test with multiple tissue classes is presented at a lower magnification in Fig. 11.7a. Predictions of the corresponding area made by both the models with the lowest and highest accuracy are compared in Fig. 11.7b and 11.7c.

11.5 Discussion and limitations

The most accurate model is the SSL based model TRI-C-SSL-AU, which improved the accuracy by 1.38% compared to the model from a pure supervised approach, TRI-SL-AF. Through a comparison of the predictions of TRI-C-SSL-AU with the other models, it also appears superior with regards to segmentation, being the model with the least faulty predictions in the annotated regions in Fig. 11.7. In addition, the prediction map in

Table 11.5: F1-Scores for each of the classes, and overall accuracy for the six models.
Green cells indicate the best result within each category.

Model	Ba	Bl	Da	Mu	St	Ur	Total
TRI-SL-AF	100.00%	98.59%	89.14%	79.42%	96.44%	98.01%	94.61%
TRI-P-SSL-F	100.00%	98.64%	90.01%	82.68%	96.14%	98.29%	95.19%
TRI-C-SSL-F	99.99%	96.66%	90.55%	82.54%	95.93%	98.59%	95.12%
TRI-SL-AU	100.00%	99.88%	87.86%	78.10%	98.10%	99.09%	94.57%
TRI-P-SSL-AU	100.00%	97.36%	88.21%	82.18%	96.79%	99.45%	94.85%
TRI-C-SSL-AU	100.00%	98.70%	91.88%	84.71%	95.92%	98.96%	95.99%

Fig. 11.7c, produced with model C-SSL-AU, appears to have less noise when compared to the others for WSI_segment_test.

Comparing the results in Table 11.5 with the different predictions in Fig. 11.6, it would be reasonable to assume that the model with the highest F1-Score for blood, TRI-SL-AU, would produce the most accurate prediction. TRI-SL-AU is trained through a traditional supervised approach on dataset $D_{gt}\{train\}$ that contains a relatively large amount of urothelium tiles, and achieves the 2nd highest F1-Score for urothelium. This is, however, quite the opposite of the situation, as it is the model that predicted the most urothelium tiles in the blood area in Fig. 11.6. This is most likely an outcome with several underlying factors: The labeled training set $D_{gt}\{train\}$ is quite small, with an even smaller test set $D_{gt}\{test\}$. It is also possible that the area in Fig. 11.6 contains features not present in the ground truth dataset.

Each WSI will typically produce hundreds of thousands of tiles, opposing a challenge when selecting tiles through a probability-based self-training method. A large number of tiles will have a high probability if the specific class is trained with many labels in the original model, i.e., more features have been learned for that class. To counter this, a minimum tile per patient threshold was set to discard WSI containing a small number of tiles, as they are most likely misclassified. This does, however, not prevent over-representation of the top-left portion of the WSIs, which will occur when a WSI contains large amounts of sufficient tiles of a certain class. One might also argue that the model will not learn that many new features from tiles it already is 100% certain about and that the method becomes more of an alternative to augmentation.

By using the cluster-based approach, it is safer to include tiles of lower probability, as it is safe to assume that tiles closer to each other are more likely to hold the same label. Also, the method ensures that tiles are distributed more evenly across the WSIs in comparison to the probability-based self-training method. This can also be seen as augmentation, and an unfrozen VGG16 model has a significant improvement when comparing the two cluster-based models TRI-C-SSL-F and TRI-C-SSL-AU, where accuracy increases from 95.12% to 95.99% respectively. The opposite effect is observed for augmenting and unfreezing with the probability-based models, decreasing the accuracy from 95.19% for TRI-P-SSL-F to 94.85% for TRI-P-SSL-AU. The SSL models without augmentation, TRI-P-SSL-F and TRI-C-SSL-F, performed relatively equal with regards to classification, however, TRI-C-SSL-F performs best with regards to segmentation.

As the models are fed three levels of magnification, where the ground truth marking is based on the 400x magnification, the corresponding 100x and 25x images contain very little of the same tissue type in some cases. This causes problems for the models, especially if the 100x and 25x images are both of a different tissue class than the 400x image. An example of this is how several tiles of ground truth label blood are predicted as background in Fig. 11.6, as this area is rather isolated from nearby tissue.

A limiting factor of this study is the small size of muscle and stroma compared to the other classes in the ground truth dataset. Augmentation techniques are implemented to try and mitigate this issue, but still, the accuracy of muscle is not as high as the other classes.

11.6 Conclusion and future work

The supervised model, TRI-SL-AF, trained only on the ground truth dataset, $D_{gt}\{train\}$, achieved an accuracy of 94.61%, with 2x augmentation of the two classes with the lowest representation. By including the cluster-weak dataset, D_{cw}, the model TRI-C-SSL-AU improved the accuracy by 1.38%. Furthermore, F1-Score stayed the same or increased for every single class, and a distinct improvement is seen when comparing the prediction maps in Fig. 11.6 and 11.7.

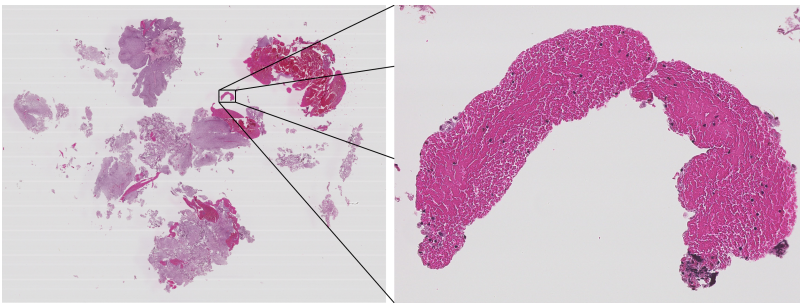
The probability-based model TRI-P-SSL-AU saw a significant improvement in classifying urothelium, with an increase of 1.44% in F1-Score, from an initial 98.08%. The accuracy was, however, only increased by 0.24%, as the model had a large reduction in F1-Score for blood.

The two different semi-supervised methods tested, both outperformed the supervised methods with regards to classification and segmentation. This shows that the combination of clusters and probability is better than only probability. The lack of labeled data makes both methods well suited to increase the training data, however, our experiments conclude that no augmentation and frozen VGG16 weights are preferred to using augmentation and unfrozen weights in a pure probability-based approach.

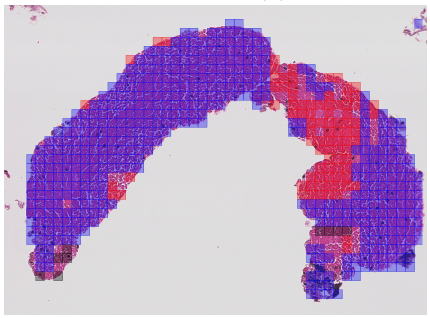
For the probability-based self-training method, better distribution of tiles in the WSI is needed for this method to be improved. This can be achieved by implementing linear spacing between all tiles of a sufficient probability score per WSI. For the cluster-based self-training method, several things can be considered for future work: a) implementing a random selection of clusters with sufficient average probability, b) selecting clusters

more evenly spaced, or c) increase criteria for stroma and muscle tissue classes. Implementing mixup [149] to generate more training data of under-represented classes could be a viable method for improving segmentation capabilities with regards to tiles of several tissue types.

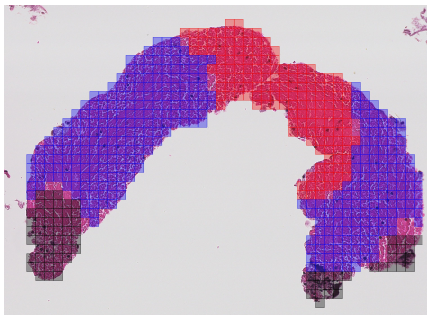
A viable segmentation method for histological images can assist pathologists in faster evaluation speeds, as pre-segmented images can immediately point out regions of interest. In addition, the system could contribute to computer-aided diagnosis systems, which can improve the rate of grading and staging of cancer and result in a more unison and objective diagnosis.



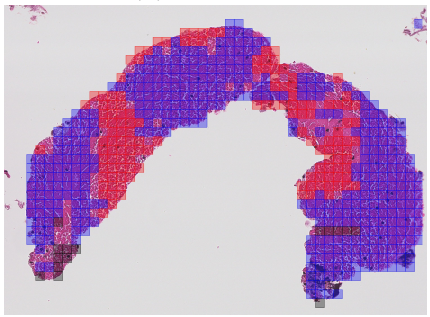
(a) Region location in WSI_segment_test.



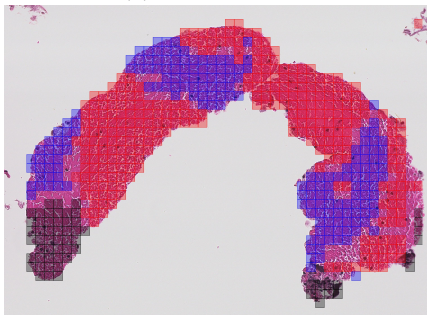
(b) TRI-SL-AF.



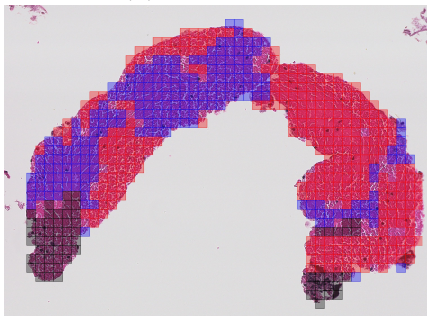
(c) TRI-SL-AU-F.



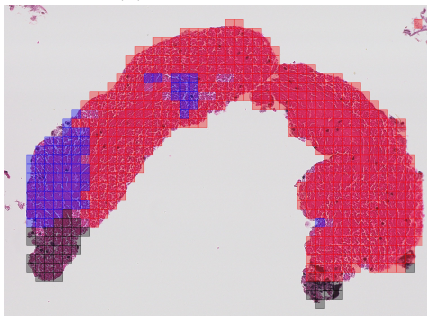
(d) TRI-P-SSL-F.



(e) TRI-P-SSL-AU.

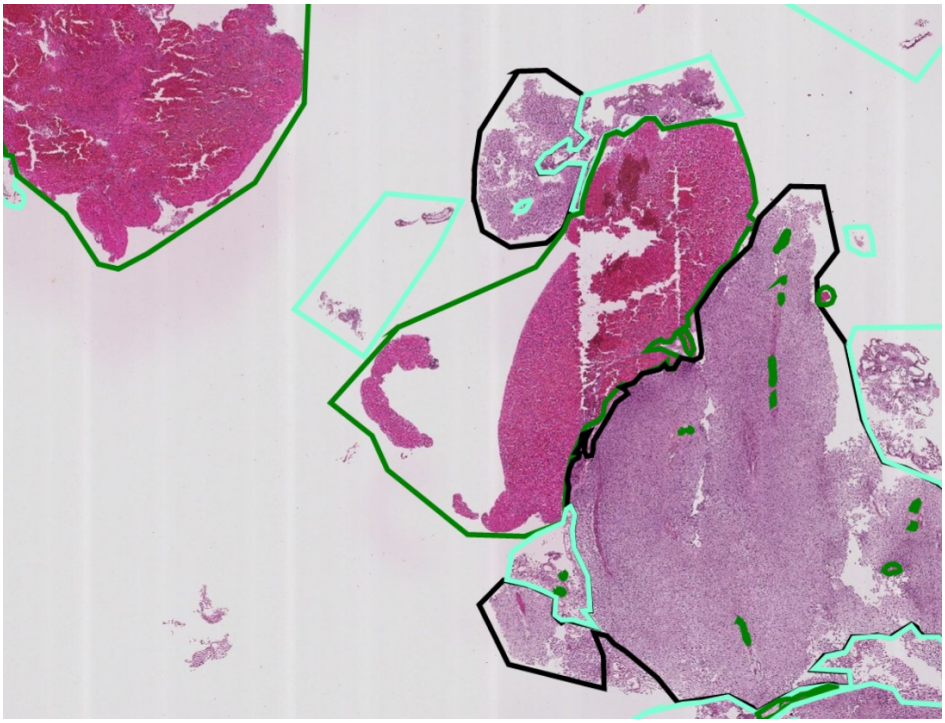


(f) TRI-C-SSL-F.

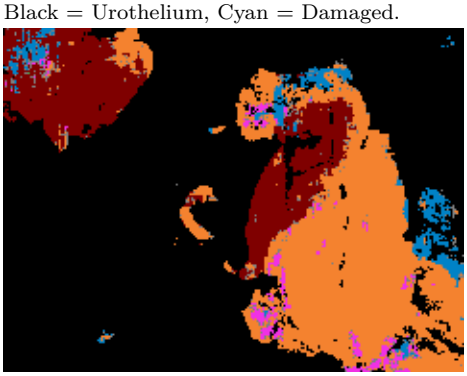


(g) TRI-C-SSL-AU.

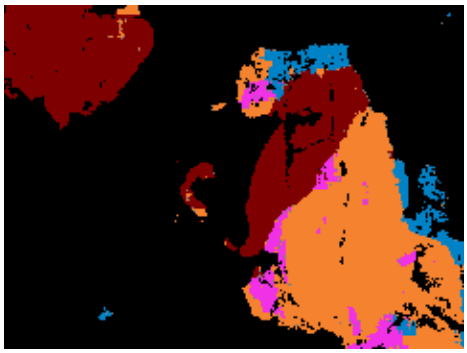
Figure 11.6: Predictions for a region in WSI_segment_test with ground truth label blood. Color specifies predicted tile class: Blue = Urothelium tissue, Red = Blood cells, Black = Background.



(a) Ground truth annotations. Colours represent ground truth annotated areas: Green = Blood, Black = Urothelium, Cyan = Damaged.



(b) TRI-SL-AF.



(c) TRI-C-SSL-AU.

Figure 11.7: Low magnification region in WSI_segment_test.
(b,c) Colours represent predicted labels: Red = Blood, Black = Background, Orange = Urothelium, Blue = Damaged, Pink = Stroma, Green = Muscle, Grey = Undefined.

Paper 4: Parameterized Extraction of Tiles in Multilevel Gigapixel Images

Parameterized Extraction of Tiles in Multilevel Gigapixel Images

Rune Wetteland[1], **Kjersti Engan**[1], **Trygve Eftestøl**[1]

[1] Department of Electrical Engineering and Computer Science, University of Stavanger, Norway

Published by IEEE in the Proceedings of the 12th International Symposium on Image and Signal Processing and Analysis (ISPA), 2021.

https://ieeexplore.ieee.org/document/9552104

Paper 5:
Automatic Diagnostic Tool for Predicting Cancer Grade in Bladder Cancer Patients Using Deep Learning

Automatic Diagnostic Tool for Predicting Cancer Grade in Bladder Cancer Patients Using Deep Learning

Rune Wetteland[1], Vebjørn Kvikstad[2,3], Trygve Eftestøl[1], Erlend Tøssebro[1], Melinda Lillesand[2], Emiel A. M. Janssen[2,3], Kjersti Engan[1]

[1] Department of Electrical Engineering and Computer Science, University of Stavanger, Norway

[2] Department of Pathology, Stavanger University Hospital, Norway

[3] Department of Chemistry, Bioscience and Environmental Engineering, University of Stavanger, Norway

Published in the Journal IEEE Access, 2021.

https://ieeexplore.ieee.org/document/9513308

Abstract:
The most common type of bladder cancer is urothelial carcinoma, which is among the cancer types with the highest recurrence rate and lifetime treatment cost per patient. Diagnosed patients are stratified into risk groups, mainly based on grade and stage. However, it is well known that correct grading of bladder cancer suffers from intra- and interobserver variability and inconsistent reproducibility between pathologists, potentially leading to under- or overtreatment of the patients. The economic burden, unnecessary patient suffering, and additional load on the health care system illustrate the importance of developing new tools to aid pathologists. We propose a pipeline, called $\text{TRI}_{\text{grade}}$, that will identify diagnostic relevant regions in the whole-slide image (WSI) and collectively predict the grade of the current WSI. The system consists of two main models, trained on weak slide-level grade labels. First, a WSI is segmented into the different tissue classes (urothelium, stroma, muscle, blood, damaged tissue, and background). Next, tiles are extracted from the diagnostic relevant urothelium tissue from three magnification levels (25x, 100x, and 400x) and processed sequentially by a convolutional neural network (CNN) based model. Ten models were trained for the slide-level grading experiment, where the best model achieved an F1-score of 0.90 on a test set consisting of 50 WSIs. The best model was further evaluated on a smaller segmentation test set, consisting of 14 WSIs where low- and high-grade regions were annotated by a pathologist. The $\text{TRI}_{\text{grade}}$ pipeline achieved an average F1-score of 0.91 for both the low-grade and high-grade classes.

13.1 Introduction

Bladder cancer is the 10^{th} most commonly diagnosed cancer disease worldwide, with 573 278 new cases in 2020 [119]. The most common type of bladder cancer is urothelial carcinoma, in which men are overrepresented. It is among the cancer types with the highest recurrence rate, approximatly 50 to 70%, which makes it especially challenging [81]. It requires an intensive treatment and follow-up plan, which results in it being one of the cancer types with the highest lifetime treatment cost per patient [12, 111]. In the case of muscle-invasive bladder cancer (MIBC), where the cancer has invaded the muscle wall of the bladder, a cystectomy is often required. However, cancers that stay confined in the bladder mucosa are referred to as non-muscle-invasive bladder cancer (NMIBC) and are easier to treat.

In histopathological diagnostics, pathologists use grading and staging to describe the tumor. These parameters are used to stratify patients into risk groups and form a suitable treatment and follow-up plan. The grade of a tumor describes the differentiation state of the tumor cells under a microscope. Different cancers have different grading scales, but in general, if the cancer cells are similar to that of healthy non-cancerous cells, the grade will be low, and the cancer will have a lower likelihood of spreading. On the other hand, if the cells have a more abnormal appearance and are disorganized, the grade will be higher. In addition to the grade, tumor stage is also important and is determined by the size of the primary tumor, how far it has spread into the surrounding tissue, and the number of primary tumors present. In this paper, we focus on grading of NMIBC. However, it is well known that correct grading of bladder cancer suffers from intra- and interobserver variability and inconsistent reproducibility between pathologists [65, 82], which can lead to both under- or overtreatment of the patients. New tools to aid pathologists are therefore desired.

The World Health Organization (WHO) has proposed three grading systems for bladder cancer. The first grading system was introduced in 1973, referred to as WHO73, which is still somewhat used today. It consists of three categories, grade 1, grade 2, and grade 3, where grade 3 is the most severe state. A revised edition of the grading system was introduced in 2004 called WHO04, and further updated in 2016 as WHO16. In these versions, cases are split into low- and high-grade carcinoma. Some examples of low- and high-grade areas are shown in Fig. 13.1. Grade 1 patients are referred to as low-grade patients, and grade 3 patients are high-grade patients. Patients diagnosed as grade 2, however, are now split into either

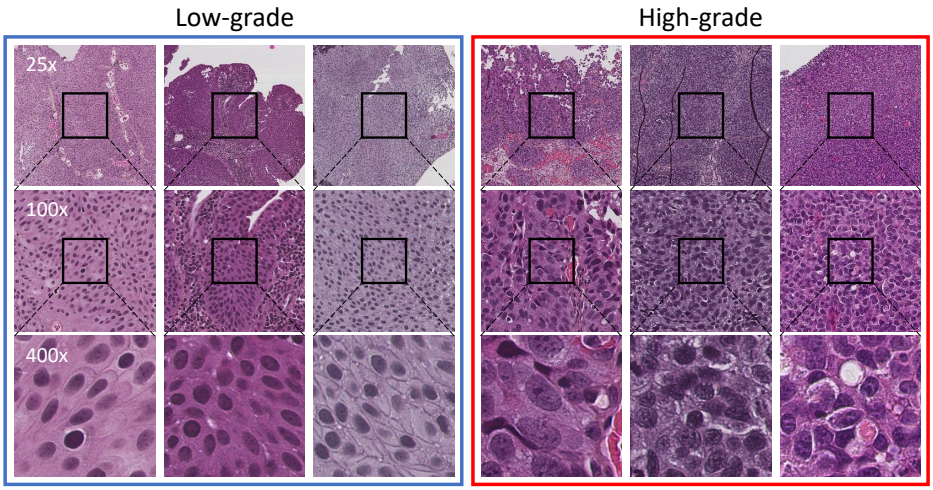


Figure 13.1: Examples of low-grade and high-grade tiles extracted from a WSI. The tiles are extracted from three magnification levels (25x, 100x, and 400x) and all have the same size of 256 × 256 pixels.

the low- or high-grade case. This might seem like a minor change, but for a patient to be diagnosed as low- or high-grade may result in very different follow-up regimes and local treatment with potential adverse events. A patient falsely diagnosed as a high-risk patient is an example of unnecessary patient suffering by overtreatment, additional load on the health care system, and extra cost. The data material used in this paper was collected and diagnosed prior to 2016 and will therefore focus on the WHO04 grading system.

After the tumor is removed, it is placed on an object glass and stained before a pathologist examines it. This is usually done through a microscope; however, with the introduction of digital pathology, digital versions of the stained specimen are also available in the form of whole-slide images (WSI). This has multiple advantages, such as remote access, storage and sharing cases between institutes, cloud computing, improved workflow, as well as computational pathology, which enables the use of new tools to process and interpret the tissue samples. All of which can improve the diagnostic accuracy and the clinical outcome of the patients [10, 16, 51, 79, 90].

Recent years have seen a rapid increase in both interest and usage of machine learning applications. Such tools could potentially be used to assist pathologists and help reduce the increasing workload. Also, because

the errors made by a machine learning system may be different from that of a pathologist, the two may be combined for improved accuracy by the pathologist, as shown by Wang et al. [133]. Low reproducibility and variability in interpretations may also be reduced if a trustworthy computer-aided diagnosis (CAD) system could be implemented in a clinical setting.

With a CAD system, we want to map a WSI input to one of the disease output categories. The traditional machine-learning method to achieve this is by supervised learning. A set of known image and label pairs are shown to the model, which uses a gradient descent algorithm to optimize its parameters. For these algorithms to work efficiently and create robust models, a large set of image-label pairs are needed. Within digital pathology, we have access to a large amount of image data in the form of WSIs. However, annotated data is limited, challenging the practicability of supervised learning approaches. The nature of the images also calls for expert input to be able to annotate them. This is a time-consuming and, in some cases, challenging task. To create enough of the image-label pairs necessary to train these models and avoid the expensive annotation process, one possibility is to utilize data already available in the form of the slide-level diagnosis information. The WSIs are split into smaller images in the form of tiles, and the slide-level diagnosis will be assigned to each of the tiles.

For patients diagnosed with NMIBC, the tumor is usually removed through transurethral resection of bladder tumor (TURBT). During this process, parts of the tissue get damaged, either heating damage from the cauterization process or physical damage from tearing. Other tissue types, like stroma or muscle, as well as blood, are also often present in the slides of urothelial carcinoma. For the purpose of grading NMIBC, urothelium is the most diagnostic relevant tissue. For staging, both urothelium and stroma, and particularly the border between them, is essential. The presence of muscle tissue also has importance for correct staging. However, cauterized tissue from the TURBT process, as well as areas containing blood, have no diagnostic relevance. Feeding a deep learning model with these irrelevant tissue classes, e.g., blood or damaged tissue, may harm the diagnostic model's accuracy. To avoid this, we have previously proposed a method based on convolutional neural networks (CNN), which automatically segments NMIBC slides into background and five foreground classes (urothelium, stroma, muscle, blood, and damaged tissue). This tissue

classification model is referred to as the $\text{TRI}_{\text{tissue}}$-model in the following and is explained in detail in Wetteland et al. [140].

In the current paper, we propose a system called $\text{TRI}_{\text{grade}}$ for automatically grading WSI according to the WHO04 grading system. The proposed system uses the $\text{TRI}_{\text{tissue}}$-model as a first-stage network for preprocessing the WSI to find regions of urothelium tissue. The extracted urothelium tissue is then fed through a second-stage network called the $\text{TRI}_{\text{WHO04}}$-model for automatic grading.

The large size of the gigapixel images causes some challenges. It is not possible to feed the entire image into a deep learning algorithm; instead, tiles of a suitable size are extracted from the WSI and fed to the algorithm sequentially. The CNN-based model assigns a prediction score to every tile. These predictions are used to create a heatmap showing which regions were predicted with low- or high-grade carcinoma. The final decision can further be aggregated from the micro predictions into a slide-level prediction.

A WSI is stored in a pyramid format with multiple magnification levels, where the different levels will give different information. An example of such a pyramidal WSI is shown in Fig. 13.2. A pathologist will typically zoom in and out of a WSI to gather information at several scales before reaching a final decision. Our proposed method mimics this behavior by combining global context information and local details by utilizing a multiscale model architecture.

13.1.1 Previous work

With the introduction of digital pathology, there has been an increase in medical application research utilizing machine learning and deep learning approaches. Most research is related to cancer diseases such as breast, lung, prostate, brain, and skin cancer [88]. By looking at the list of US Food & Drugs Administration (FDA) approved artificial intelligence (AI) based medical technologies, most are in the fields of radiology, cardiology, and Internal Medicine/General Practice [9]. Still, a lot of effort is also aimed towards histological images [19, 25, 45, 57, 117].

The majority of CAD research conducted on histological images utilize two or more seperate models in their methods [19, 59, 78, 115, 151]. First, a segmentation algorithm or region of interest (ROI) selection step is performed to narrow down the area which needs additional processing. This is an important step that helps in several ways. Compared to standard

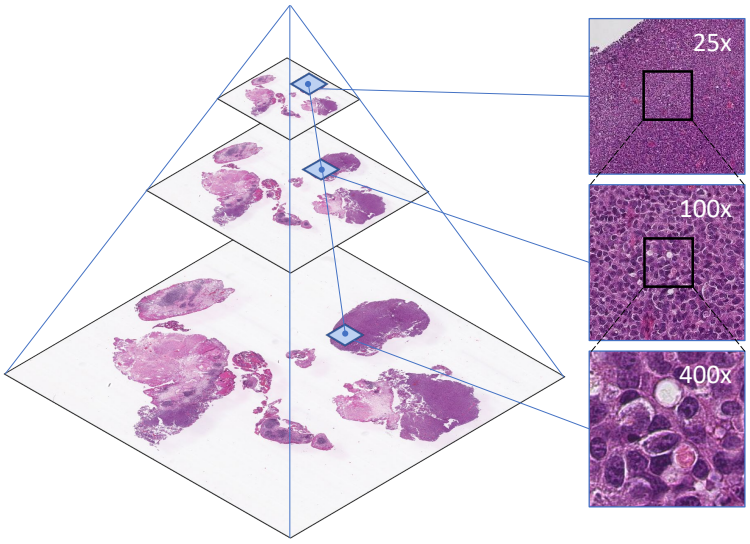


Figure 13.2: WSI images are stored in a pyramidal format, where the base image corresponds to the highest magnification level. The right-hand side shows a set of three tiles extracted so that the center of the tile corresponds to the same physical area in the WSI, forming a triplet.

images, the WSIs are very large in size, and it is computationally expensive to process the entire WSI. By limiting the number of extracted tiles, the classification runtime is reduced, speeding up the classification step. Also, by removing the unwanted and diagnostically irrelevant areas, the extracted datasets will consist of higher quality tiles, which aids the classification algorithm in the following steps. After segmentation, tiles from the ROI are processed, usually by a classification model, which will predict the class of the tiles. Examples of tile classes can be cancer vs. non-cancer, recurrence vs. no recurrence, cancer grading or staging, or other classes related to cancer diagnosis. After all the selected tiles have been classified, the predictions are aggregated into a final slide-level prediction, usually using statistical or machine-learning methods.

Some research has been aimed towards urothelial carcinoma, otherwise known as bladder cancer. In Jansen et al. [59], they utilized two individual single-scale neural networks to detect and grade 328 cases of bladder cancer collected from 232 patients. A U-net-based segmentation network was trained to detect and segment the urothelium tissue, used as input to a second network trained to grade the urothelium tissue according

Table 13.1: Overview of how the data material in this study is distributed into training, validation, and test sets. For triplets in the training dataset, see Table 13.2.

	Low-grade WSIs	High-grade WSIs	Total WSIs	Total triplets
Training	124	96	220	Table 13.2
Validation	17	13	30	301 775
Test	28	22	50	473 678

to the WHO04 grading system. The classification network assessed the WHO04 grading on slide-level, using the majority vote of all classified tiles. The predictions were compared with the grading of three experienced pathologists. According to the consensus reading, the classification model achieved an accuracy score of 74%. The included whole-slide images were all exported at 20x magnification (0.5 µm per pixel).

From the same research group, the work of Lucas et al. [78] utilized the same urothelium segmentation model as presented in [59]. Regions of urothelium were then fed into a selection network which classified tiles into recurrence vs. no recurrence. A strategy was applied to select features from 200 tiles fed into a final bidirectional gated recurrent unit (GRU) classification network that predicts 1-year and 5-year recurrence-free survival (RFS) in bladder cancer patients.

The work of Zhang et al. [151] was also performed on bladder cancer. They used three different neural networks referred to as s-net, d-net, and a-net. The s-net model is a U-net-like architecture that classifies each pixel as tumor vs. non-tumor. The d-net then characterizes the tumor ROIs and generates an interpretable diagnosis and low-dimensional encodings. Finally, the a-net uses the ROI encodings and predicts a slide-level WHO04 grading.

Multiscale cancer subtype classification, where two or more different magnification scales are fed to the classification model, has been shown to improve the accuracy compared to single-scale models [114, 140]. This mimics the pathologist's process, which will zoom in and out to investigate the tissue area at several scales.

In Skrede et al. [115] the WSI is first segmented, before tiles are extracted at 10x and 40x resolution. The tiles from each scale are fed to an ensemble of 5 models, using a total of ten CNN-based models. The average score from the ensembles is used to predict the prognosis of colorectal patients.

Table 13.2: Extracted triplets for the training dataset.

N	Total triplets before aug.	Total triplets after aug.	Percentage increase
250	54 564	55 000	0.8%
500	106 577	110 000	3.1%
1 000	202 904	219 560	7.6%
3 000	534 734	647 368	17.4%
5 000	812 588	1 051 752	22.7%

In the work of Hashimoto et al. [52] WSIs from malignant lymphoma were fed to a multiscale CNN-based model. They compared the results of models using tiles extracted at 10x or 20x resolution. However, the best result was achieved by combining the two scales into a multiscale model. The authors of this study also confirm that class-specific features exist at different magnification scales.

Previous work from our group, on bladder cancer, included tissue segmentation [32, 138, 140], and prediction of recurrence in NMIBC patients [127]. In Wetteland et al. [140], we experimented with three magnification scales and any combination of these. We proposed three MONO-models (Mono-25x, Mono-100x, and Mono-400x), three DI-models (DI-25x-100x, DI-25x-400x, and DI-100x-400x), and finally a model utilizing all three magnification scales, TRI-25x-100x-400x. All models used the VGG16 network as a feature extractor and were trained and evaluated on six tissue classes. The MONO-models performed worst, and the best result was achieved with the TRI-model utilizing all scales, supporting the claim that multiscale models achieve better results. Both frozen and unfrozen weights were experimented with, but the TRI-model trained with frozen weights in the VGG16 models performed best and achieved an average F1-score of 96.5% when evaluated on all six classes, and an average F1-score of 97.6% for the urothelium class alone.

Based on this result, we continued with the TRI-model and VGG16 as feature extractors in the current paper. We have not evaluated the MONO- or DI-models on the diagnostic data. The model referred to as TRI-25x-100x-400x in [140] is in the current paper referred to as the $\text{TRI}_{\text{tissue}}$-model. It is used for tissue extraction as shown in Fig. 13.4. The name, architecture, and base model have also been carried over to this

paper and are the basis for the $\text{TRI}_{\text{WHO04}}$-model we propose here.

13.1.2 Our contributions

The current study's main contributions is listed below.

- A novel, fully automated pipeline called $\text{TRI}_{\text{grade}}$ is proposed. The system consists of a tissue segmentation model and a diagnostic WHO04 grade model. The system's output consists of a tissue segmentation map, a WHO04 heatmap, and a predicted slide-level WHO04 grade. The proposed $\text{TRI}_{\text{grade}}$ system correctly predicted 45 of the 50 WSIs in the test set, achieving an accuracy of 90%.
- The $\text{TRI}_{\text{grade}}$ system-generated heatmaps are both visualized and evaluated against a segmentation test set. This helps to demonstrate the usage of such a system for a pathologist in a clinical setting.
- An algorithm for finding the optimal value of a decision threshold for classifying WSIs at slide-level is proposed.
- We trained models on differently sized training sets. The results for this provide insight on how dataset sizes affect the performance of the models, training time per epoch, and trained epochs before reaching stopping criteria during early stopping.
- Source code for this paper is accessible at the following URL address `https://git.io/J3OdW`.

13.2 Methods

The proposed $\text{TRI}_{\text{grade}}$ system presented in this paper utilizes multiscale models, which use tiles extracted at multiple magnification levels as input. For improved readability, we define these tiles as a *triplet*. A triplet is denoted T_i and is defined as a set of three tiles extracted from a WSI at three different magnification levels (25x, 100x, and 400x). Let $\mathcal{T}$ denote a set of triplets in a WSI, where $\mathcal{T} = \{T_1, T_2 \dots \ T_i \dots \ T_{max}\}$, and the number of elements in the set is given by the cardinality $|\mathcal{T}|$. An example of a triplet is shown in Fig. 13.2.

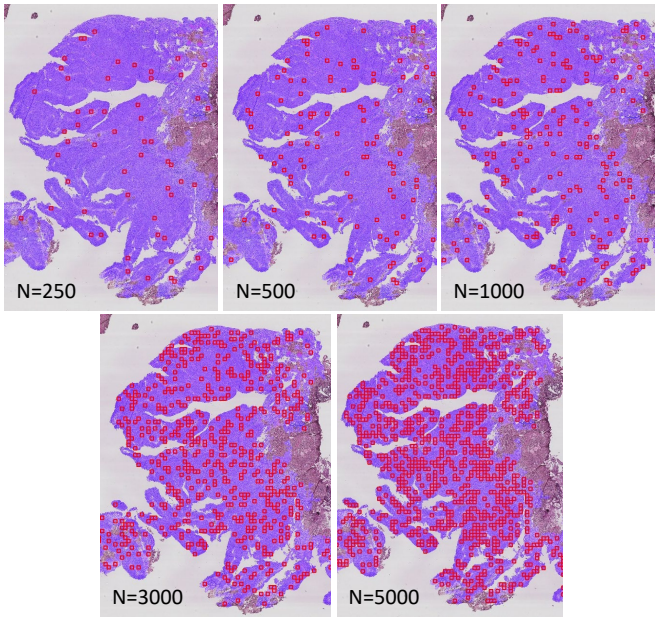


Figure 13.3: A close-up image from a WSI with a superimposed urothelium ROI mask (semi-purple). As N increases, the density of the tiles (red squares) also increases. The illustrated tiles are shown on 400x magnification level, but tiles from 25x and 100x are also extracted.

13.2.1 Data material

The data material consists of 300 digital whole-slide images from patients diagnosed with NMIBC, where the tissue is removed from the patient through transurethral resection of bladder tumor. The data was collected at the University Hospital of Stavanger, Norway, in the period 2002-2011. All non-muscle invasive bladder cancers are included in the dataset, making it a true population based dataset. The biopsies were formalin-fixed and paraffin-embedded, from which 4 µm thick sections were cut and stained with Hematoxylin, Eosin, and Saffron (HES).

The slides were diagnosed and graded according to WHO73 and WHO04 [7]. All slides have the label low-grade or high-grade in the WHO04 system. In addition, cancer stage and follow-up data on recurrence and disease progression are recorded, and all patients have stage Ta or T1, i.e., non-muscle invasive. We have, however, no annotated regions with healthy non-cancerous urothelium available. All WSI have gone through a manual quality check at the department of pathology, Stavanger University Hospital,

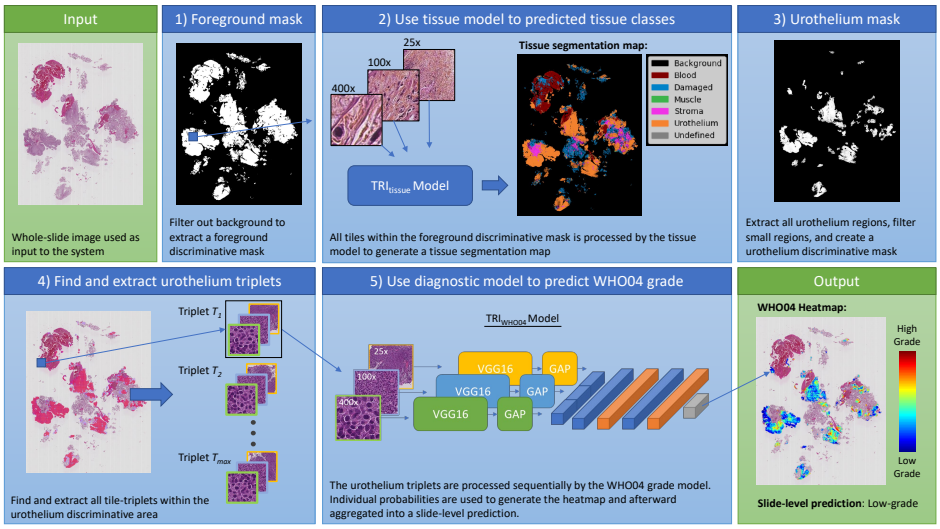


Figure 13.4: This figure presents the pipeline for our proposed system, $\text{TRI}_{\text{grade}}$. **Input)** A WSI of urothelial carcinoma is used as input. **1)** A foreground discriminative mask is found by evaluating the pixel intensity values and used as a reference to extract tiles from the WSI. **2)** The $\text{TRI}_{\text{tissue}}$-model is used to generate a tissue segmentation map. **3)** The urothelium regions are used to create a urothelium discriminative mask. **4)** Using the urothelium mask, triplets consisting of tiles from three magnification levels are extracted from the input WSI. **5)** The urothelium triplets are fed sequentially to the $\text{TRI}_{\text{WHO04}}$-model, which outputs a probabilistic score for the two classes, low- and high-grade carcinoma. **Output)** The system will output a WHO04 grade heatmap and a slide-level WHO04 prediction.

and only high-quality slides, with little or no blur, have been included in the dataset. However, as mentioned, NMIBC is removed by cauterization, which will leave burned and damaged tissue areas. All WSI are from the same laboratory, and the variation in staining color is relatively low. Ethical approval from Regional Committees for Medical and Health Research Ethics (REC), Norway, ref.no.: 2011/1539, regulated according to the Norwegian Health Research Act.

The glass slides were digitized using a Leica SCN400 slide scanner, producing WSI images in the vendor-specific scn file format. These WSI images are gigapixel images with a typical resolution of $100\,000 \times 100\,000$ pixels, stored as a pyramidal tiled image with several down-sampled versions of the base image in the same file to accommodate for rapid panning and zooming. The pyramidal structure of the WSI is depicted in Fig. 13.2. The Vips library [84] can extract the base image and the down-sampled

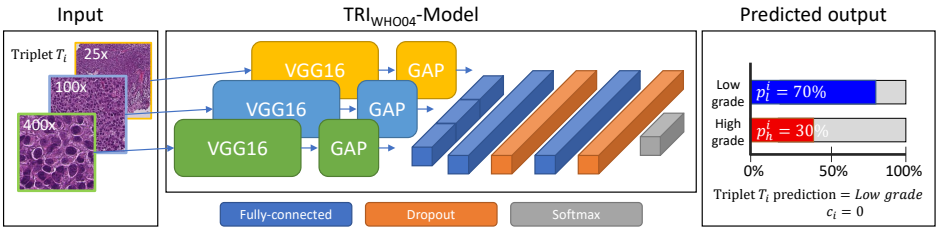


Figure 13.5: Architecture of the TRI$_{WHO04}$-model. Three separate VGG16 networks are used to extract features from each magnification scale. The global average pooling layer (GAP) is used to flatten the features into feature vectors, which are concatenated. The classification network consists of fully-connected layers and dropout layers. The output uses a softmax activation function to predict the input tiles to the two classes, low-grade and high-grade carcinoma.

versions, making it easy to extract the dataset at each resolution.

Tiles are extracted from the image pyramid at levels corresponding to 25x, 100x and 400x magnification, which is equivalent to a spatial resolution of 4 µm/pixel, 1 µm/pixel and 0.25 µm/pixel, respectively. For the TRI$_{tissue}$-model, we used a tile size of 128 × 128 pixels, which for the three magnification levels correspond to (512 µm × 512 µm), (128 µm × 128 µm), and (32 µm × 32 µm). For the TRI$_{WHO04}$-model, we had access to a much larger library of WSIs, and thus a larger tile size of 256 × 256 pixels was chosen. For the three magnification levels, this corresponds to (1 024 µm × 1 024 µm), (256 µm × 256 µm), and (64 µm × 64 µm).

The 300 WSIs included in this study were split into 220/30/50 WSIs for training, validation, and testing, respectively. Demographic characteristics of the data material were not used when splitting the data material into the different datasets. Instead, the WSIs were randomly selected and stratified to include the same ratio of all diagnostic outcomes based on the WHO73 and WHO04 grading, stage, recurrence, and disease progression, to represent the data material best. The distribution of low- and high-grade WSIs in each dataset, as well as the number of triplets in the validation and test set, can be seen in Table 13.1.

The 50 WSIs in the test set will use the slide-level label as ground truth to evaluate the TRI$_{WHO04}$-model. In addition, a pathologist has carefully annotated low- and high-grade regions in 14 of the 50 WSIs. The 14 WSIs are a sub-set of the test set and are referred to as the *segmentation test set* and will be used to evaluate the low- and high-grade segmentation performance of the best TRI$_{WHO04}$-model.

From the 220 WSIs used for training, five datasets were extracted with a different number of triplets extracted from each WSI. A set of N triplets was selected randomly from the predicted urothelium regions in each WSI, where N was set to 250, 500, 1 000, 3 000, and 5 000.

Some of the WSIs in the data material contain only small amounts of urothelium, either because the tissue sample itself is small or because most of the tissue sample consists of damaged tissue or other tissue classes. For these WSIs, an augmentation strategy was employed, where a randomly selected set of triplets were augmented. The aim of this process is for each WSI to contribute equally, or as close as possible, to the number of triplets specified by N. Augmentation was performed by rotation and vertical/horizontal mirroring of the individual tiles in the triplet. All tiles in the triplet were augmented in the same manner. By combining rotation and mirroring, a tile can be oriented in eight uniquely defined ways, making this the maximum number a particular tile can be augmented. For $N \geq 1\,000$, some WSIs did not reach the desired number of triplets, even with 8x augmentation. No augmentation was performed on the validation or test datasets. Table 13.2 shows a list of total triplets extracted, before and after augmentation, for each value of N.

Fig. 13.3 shows a region from one WSI with the extracted tiles superimposed. The semi-transparent purple color indicates the predicted urothelium region. From this region, N randomly selected tiles are extracted as indicated by the red tiles on the image. As N increase, the density of extracted tiles also increases. Also, note that only the tile extracted at magnification level 400x is visualized in the figure. At each tile position, tiles from all three magnification levels (25x, 100x, and 400x) are extracted in such a manner that the center position of each tile corresponds to the same physical location, as illustrated in Fig. 13.2.

For preprocessing, all pixel intensity values were normalized from 0-255 values into 0-1 values, and the order of the color channels was altered from RGB to BGR. These steps ensure that the input data is presented to the VGG16 network in the same fashion as when it was pre-trained on the ImageNet data. No stain normalization was performed on the extracted tiles.

Our data material contains slide-level diagnostic information; however, no location annotations exist, showing where in the WSI the low- or high-grade regions are found, except on our segmentation test set, as explained. As manual annotation is time-consuming, expensive, and requires expert input,

it is not feasible to get this type of detailed annotations on large datasets as needed for training such models, particularly considering both the size of each WSI and the total number of WSIs in the data material. Instead, each extracted tile inherits the slide-level WHO04 grade as its label. This is not ideal, as high-grade slides may contain regions with low-grade tissue. Consequently, all the extracted datasets are thus regarded as weakly labeled due to the inaccurate labels, which is consistent with what is called a weak label in many tasks [27]. The segmentation test set is considered strongly labeled.

13.2.2 Proposed system

We propose a pipeline, called $\text{TRI}_{\text{grade}}$, that takes a WSI as input and outputs a tissue segmentation map, a WHO04 grading heatmap, and a slide-level WHO04 grade prediction. The pipeline consists of two main models, denoted as $\text{TRI}_{\text{tissue}}$-model and $\text{TRI}_{\text{WHO04}}$-model. The task of the $\text{TRI}_{\text{tissue}}$-model is to classify an input triplet as a tissue type which then can be used to make a tissue segmentation map. The task of the $\text{TRI}_{\text{WHO04}}$-model is predicting the cancer grade, i.e., low- or high-grade, based on the urothelium tissue. The $\text{TRI}_{\text{grade}}$ pipeline is depicted in Fig. 13.4 and explained in detail below.

$\text{TRI}_{\text{grade}}$ Pipeline

The $\text{TRI}_{\text{grade}}$ pipeline depicted in Fig. 13.4 contains five steps explained here. The input to the pipeline consists of a WSI file in the vendor-specific .scn file format. First, in step 1, a foreground discriminative mask is found on the 400x level by evaluating the pixel intensity values as grey background or not. Using the foreground mask as reference, tiles with dimension 128 $\times$ 128 pixels were extracted from the WSI with 87.5% overlap, resulting in the inner 16 $\times$ 16 pixels being classified for each tile. Three tiles were extracted in the WSI (25x, 100x, and 400x) for each location, forming a triplet. All tiles in each triplet have the same dimension of 128 $\times$ 128 pixels and are extracted such as the center point corresponds to the same physical location in the WSI for all three tiles, as shown in Fig. 13.2.

In step 2, triplets are sequentially fed into the $\text{TRI}_{\text{tissue}}$-model we proposed in Wetteland et al. [140]. This model will evaluate the triplets and predict which of the six tissue classes (urothelium, stroma, muscle, blood, damaged tissue, and background) the current triplet belongs. In our case, the class

of damaged tissue is a collection of all tissue that is not one of the other classes, and in our dataset, this is mainly cauterized or torn tissue areas. If blurred regions are a problem in the dataset, this can be made as a separate class or included in the damaged tissue class. After predicting all triplets, a segmented tissue map is created, visualizing all tissue regions in the WSI. This tissue map can also be presented to the clinician to help guide them more efficiently to the specific tissue types in the WSI.

From the generated tissue map, all urothelium regions are extracted in step 3. Small regions are filtered to suppress noise, and a urothelium discriminative mask is created on the 400x level. In step 4, a grid of non-overlapping tiles is overlayed on the WSI at the 400x level, this time using tiles of dimension 256×256 pixels. The individual tiles in the grid are checked against the discrimination mask. If 80% or more of a tile lay within the discriminate mask, the position is saved, while the remaining tiles are discarded. For the validation and test sets, triplets from all the saved positions are extracted. Whereas for the training set, N randomly selected triplets are extracted from the saved positions, where training sets are formed with N set to 250, 500, 1 000, 3 000, and 5 000. If fewer than N positions are saved, the augmentation strategy explained in the data material section is employed. The total number of extracted triplets for each dataset is shown in Tables 13.1 and 13.2.

A comprehensive description of how triplets are extracted from the WSI is given in Wetteland et al. [137], where a parameterized method for extracting tiles in multilevel images is given. The parameters used in this paper are the tile size parameter $L_T = 256$. The overlap-ratio between a tile and the discriminative mask is set to 80%, which corresponds to a value of $\phi = 0.8$. Tiles are checked at the 400x level by setting $\alpha = 0$, and the corresponding tiles in the triplets are found at level 25x and 100x, i.e., $\mathcal{S}_\beta = \{1, 2\}$. The binary mask B^k is set as the urothelium discriminative mask, and the starting coordinate of the grid is at position $(0, 0)$. With these parameters and the methods described in [137], extraction of the triplets in the WSIs is repeatable and reproducible.

In step 5, the extracted urothelium triplets are fed to the $\text{TRI}_{\text{WHO04}}$-model, which outputs a probabilistic score for the two classes, low- and high-grade carcinoma. Finally, all scores are used to generate a heatmap which is overlayed on the WSI, and the aggregated micro-predictions are measured against the decision threshold D_t to get the final slide-level prediction.

Model architectures

The proposed pipeline in Fig. 13.4 contains two CNN-based models used for different tasks; the $\text{TRI}_{\text{tissue}}$-model is used for tissue classification and the $\text{TRI}_{\text{WHO04}}$-model for grading of urothelium tissue. The models are built upon the same architecture but have different inputs and outputs. The architecture consists of three separate VGG16 networks, one for each input scale. Both the model architecture and the TRI-terminology comes from our previous work on the tissue model in Wetteland et al. [140].

The input to the $\text{TRI}_{\text{tissue}}$-model is a triplet consisting of three 128 $\times$ 128 pixel tiles (25x, 100x, and 400x). The model can predict triplets extracted from anywhere in the WSI, but a foreground discriminative mask is usually used to save processing time by removing the background. The output of the $\text{TRI}_{\text{tissue}}$-model is a probability distribution over the six predicted classes (urothelium, stroma, muscle, blood, damaged tissue, and background). The input to the $\text{TRI}_{\text{WHO04}}$-model is a triplet consisting of three 256 $\times$ 256 pixel tiles (25x, 100x, and 400x) extracted from urothelium tissue regions. The model outputs a probability distribution over the two predicted classes, low- and high-grade carcinoma. A block diagram of the $\text{TRI}_{\text{WHO04}}$-model architecture is depicted in Fig. 13.5. The $\text{TRI}_{\text{tissue}}$-model has almost the same architecture but has six output classes instead of two.

The individual tiles in the input triplet are fed to separate VGG16 networks. The VGG16 networks are used as base models with weights pre-trained on the ImageNet dataset, a large dataset containing annotated photographs used for computer vision research. Each VGG16 network acts as a feature extractor and takes a high dimensional tile as input (128×128×3 or 256×256×3 pixels) and compresses it down to a feature volume (8×8×512). A global average pooling (GAP) layer is used as the output layer for each VGG16 network, transforming the feature volume into a feature vector of length 512. The three feature vectors, one for each scale, are concatenated into one final feature vector of length 1 536 and fed to the classification network.

The classification network consists of two fully-connected (FC) layers using a rectified linear unit (ReLU) activation function, each followed by a dropout layer for regularisation. Lastly, an output layer with a softmax activation function is used to provide the prediction of the model. The two FC-layers and the two dropout layers each have a dimension of 4 096 neurons, and the output layer has one output neuron for each class. The

$\text{TRI}_{\text{WHO04}}$-model consists of 67M parameters, where 23M of the parameters are trainable parameters belonging to the classification network.

Tile-level prediction

When a triplet T_i is fed to the $\text{TRI}_{\text{WHO04}}$-model, the model outputs a list of probabilities for the two classes, low- and high-grade. These probabilities are denoted as $[p_l^i, p_h^i]$. To find the class with the largest predicted probability, the argmax function is used.

$$c_i = argmax([p_l^i, p_h^i]) \tag{13.1}$$

Where c_i is the index to the predicted class for the triplet at position T_i. The low-grade class has an index of 0, and the high-grade class has an index of 1.

The proposed system can also produce a heatmap from the individual triplet probabilities, which indicates the location of low- and high-grade regions. This is useful for pathologists who can focus their limited per-patient investigation time on the diagnostic relevant areas in the WSI. A color mapping function converts the high-grade probability p_h^i into a color based on its value. This color is then superimposed on the WSI at the current triplet's position, covering the same area as the 400x magnification tile in the triplet. This results in the heatmap, as seen in the bottom-right of Fig. 13.4. Only the model's probabilistic score for the high-grade class is used to generate the heatmaps. However, because there are only two classes, a low probabilistic score of the high-grade class implicitly means a high score for the low-grade class. I.e., red highlighted regions in the heatmaps are associated with the high-grade class, and blue highlights indicate the low-grade class.

Slide-level prediction

In addition to predicting the individual triplets, we also output a WHO04 slide-level prediction. A pathologist will often assign the worst case to a slide during a clinical examination, meaning that if a high-grade region exists in the WSI, the WHO04 grading should be high-grade. However, we must assume some misclassification in the WSI from both the $\text{TRI}_{\text{tissue}}$-model and $\text{TRI}_{\text{WHO04}}$-model, so there must be a minimum amount of high-grade triplets before the slide-level prediction becomes high-grade, and we would

like to find a decision threshold, D_t, which maximizes correct prediction of the WSIs.

By summing over c_i, the number of triplets predicted as high-grade is counted, since triplets predicted as low-grade is at index 0 and thus not adding to the sum. By dividing by the total number of triplets in the WSI, we get the ratio of high-grade triplets referred to as R_{high} in this paper:

$$R_{high} = \frac{1}{|\mathcal{T}|} \sum_{i=1}^{|\mathcal{T}|} c_i \tag{13.2}$$

If R_{high} exceeds the decision threshold D_t, the slide is given the slide-level prediction of high-grade; else, it is considered low-grade.

$$\hat{Y} = \begin{cases} \text{High-grade}, & \text{if } R_{high} \geq D_t \\ \text{Low-grade}, & \text{otherwise} \end{cases} \tag{13.3}$$

Algorithm 1 describes how to find the optimal threshold value D_t. Y is considered the ground truth grading of a slide and consists of a single value, whereas $\mathcal{Y}$ is a list of all the ground truth values. The same holds for $\hat{Y}$ and $\hat{\mathcal{Y}}$, which holds a single slide-level prediction and a list of all the predictions, respectively. First, all WSIs are processed, and the ratio R_{high} for each WSI is appended to the list $\mathcal{R}$. The true grade Y of each WSI is also saved in the list $\mathcal{Y}$. All WSIs in the dataset are processed before proceeding to the next step. A set of candidate threshold values, D_c, between 0-50% are tested one at a time. For each candidate threshold, the slide-level prediction $\hat{Y}$ for all WSIs is saved to the list $\hat{\mathcal{Y}}$. The total accuracy score is then calculated for the dataset. The decision threshold D_t is chosen as the candidate threshold with the highest score, or, if more than one D_c value yielded the same maximum result, the average integer value is selected as the decision threshold D_t.

Training parameters

The $\text{TRI}_{\text{WHO04}}$-model was trained using a stochastic gradient descent (SGD) optimizer with a learning rate of $1{\times}10^{-3}$, learning rate decay of $1{\times}10^{-6}$, and momentum set to 0.9. The batch size used during training was set to 128. Both dropout layers had a dropout rate of 0.5. The cross-entropy loss function was used to optimize the model during training. The pre-trained

Algorithm 1: Find optimal threshold value D_t

Initialize: $\mathcal{Y}$, $\hat{\mathcal{Y}}$, $\mathcal{R}$, $\mathcal{D}_{c_{best}}$ are empty lists
Initialize: $Acc_{max} = 0$
for $WSI \leftarrow$ training set **do**
  Feed WSI through pipeline in Fig. 13.4
  $R_{high} = \frac{1}{|\mathcal{T}|} \sum_{i=1}^{|\mathcal{T}|} c_i$
  Append R_{high} to the list $\mathcal{R}$
  Append the true grade Y of WSI to the list $\mathcal{Y}$
end for
for $D_c \leftarrow 0$ to 50 **do**
  for $R_{high} \leftarrow \mathcal{R}$ **do**

$$\hat{Y} = \begin{cases} \text{High-grade}, & \text{if } R_{high} \geq D_c \\ \text{Low-grade}, & \text{otherwise} \end{cases}$$

    Append the slide-level prediction $\hat{Y}$ to the list $\hat{\mathcal{Y}}$
  end for
  $Acc_{D_c} =$ sklearn.metrics.accuracy_score($\mathcal{Y}$, $\hat{\mathcal{Y}}$)
  if $Acc_{D_c} > Acc_{max}$ **then**
    $Acc_{max} \leftarrow Acc_{D_c}$
    Clear list $\mathcal{D}_{c_{best}}$
  end if
  if $Acc_{D_c} \geq Acc_{max}$ **then**
    Append D_c to list $\mathcal{D}_{c_{best}}$
  end if
end for
$D_t = \lceil \frac{1}{|\mathcal{D}_{c_{best}}|} \sum \mathcal{D}_{c_{best}} \rceil$

weights of the VGG16 networks were held frozen during training. To avoid overfitting the models on the training set, an early-stopping rule monitored the validation loss and stopped the training when no improvements were seen for ten epochs. The best epoch was restored when testing the models on the test set.

To train the models, a program was written in Python 3.6 using Keras 2.2.4 together with the Tensorflow 1.14 as backend [1, 29]. The PyVips 2.1 library was used for handling the WSI [84], and Scikit-learn 0.19 for evaluation [98]. The models were training on a Ubuntu 18.04 server, running on dual Xeon E5-2650 v5 @ 2.2GHz with a total of 48 cores. An Nvidia

Tesla P100 16GB GPU was used for the training. Training parameters for the $\text{TRI}_{\text{tissue}}$-model can be found in Wetteland et al. [140].

13.3 Experiments

We have conducted two experiments, listed here.

Experiment 1: is for slide-level prediction of WHO04 grade and is tested on the test set of 50 WSIs. As training of the $\text{TRI}_{\text{WHO04}}$-model is very time-consuming, we wanted to see if it is preferable to utilize more of the available urothelium data from each WSI as training data at the cost of additional training time or if a smaller dataset could perform equally well. This is interesting, both for our research group as well as other researchers working with large WSI datasets. If the optimal number of tiles used from each WSI during training can be lowered, then time can be saved in future experiments. To investigate this, we created several datasets where we extracted N triplets per WSI, as shown in Table 13.2. In this experiment, ten versions of the $\text{TRI}_{\text{WHO04}}$-model, all with the same architecture, were trained on training sets of various sizes, listed in Table 13.2. The micro predictions from the individual triplets were aggregated into a slide-level prediction of the WHO04 grading. A decision threshold D_t was found for each model using Algorithm 1; then, equation 13.3 was used to provide the final predicted grade.

Experiment 2: is testing the tile-level prediction and compare that in detail with the 14 WSIs of the segmentation test set. This set contains pathologist annotated regions belonging to either low- or high-grade which are considered the ground truth. The best model from experiment 1 is used for this, and the model's performance will be visualized as heatmaps. Calculation of recall and F1-score will be presented for each WSI, in addition to a total score across all WSIs.

13.4 Results

In experiment 1, slide-level test results for the ten models are listed in Table 13.3, showing trained epochs, time, precision, recall, F1-score, and the threshold value D_t. For precision, recall, and F1-score, the weighted average score is presented as reported by the *classification report* function from the scikit-learn library [98].

Table 13.3: Slide-level prediction results for automatic WHO04 grading tested on the 50 WSIs of the test set. Precision, recall, and F1-score is the weighted average score for the two classes across all 50 WSIs in the test set. D_t is the decision threshold found using Algorithm 1. The column trained epochs show how many epochs each model was trained before the early stopping criteria were reached. Training times are shown as hours:minutes.

Model	Trained epochs	Time per epoch	Training time	Precision	Recall	F1-Score	D_t
$\text{TRI}_{\text{WHO04}}$-250	23	1:22	31:39	0.86	0.84	0.84	49
$\text{TRI}_{\text{WHO04}}$-250-AUG	33	1:19	43:53	0.89	0.86	0.85	47
$\text{TRI}_{\text{WHO04}}$-500	21	1:42	35:59	0.89	0.86	0.85	43
$\text{TRI}_{\text{WHO04}}$-500-AUG	21	1:44	36:39	0.77	0.76	0.76	49
$\text{TRI}_{\text{WHO04}}$-1000	15	1:52	28:03	0.83	0.82	0.82	49
$\text{TRI}_{\text{WHO04}}$-1000-AUG	18	1:55	34:45	0.80	0.80	0.80	49
$\text{TRI}_{\text{WHO04}}$-3000	15	3:24	51:01	0.89	0.86	0.85	49
$\text{TRI}_{\text{WHO04}}$-3000-AUG	12	3:29	41:54	0.78	0.78	0.78	49
$\text{TRI}_{\text{WHO04}}$-5000	16	4:10	66:42	0.85	0.84	0.84	48
$\text{TRI}_{\text{WHO04}}$-5000-AUG	17	5:18	90:20	**0.92**	**0.90**	**0.90**	48

For experiment 2, the $\text{TRI}_{\text{WHO04}}$-5000-AUG model was used, as it performed best in experiment 1. The predicted heatmaps for each WSI in the segmentation test set are shown in Fig. 13.6 together with the ground truth. Recall, and F1-score for each WSI is listed in Table 13.4. As each ground truth WSIs only contain annotations for one of the two classes, the precision score will always be 1.00 because whenever the model predicts the ground truth class, it will be correct. The precision column in Table 13.4 is thus discarded. The last row in Table 13.4 shows the average value of all scores for each class together with the standard deviation. Table 13.5 shows the total aggregated results for all 14 WSIs. Here, the predictions for all WSIs are accumulated before the score is calculated.

A slide-level comparison between the proposed $\text{TRI}_{\text{grade}}$ system and the model presented in Jansen et al. [59] is shown in Table 13.6. The $\text{TRI}_{\text{grade}}$ system consists of the $\text{TRI}_{\text{tissue}}$-model followed by the $\text{TRI}_{\text{WHO04}}$-5000-AUG model. Values for sensitivity, specificity, and accuracy are shown for easier comparison with the reported results from [59]. These values are unweighted and calculated using values for true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Note that these results are based on models trained and evaluated on different datasets.

13.5 Discussion

The three VGG16 networks are identical copies as we have used frozen (pre-trained) weights in this work. Thus, it would be possible to use only one copy of the model, with the appropriate change in the architecture, keeping in mind that the feature vectors from the different magnifications are concatenated before the classification network. However, utilizing three versions of the VGG16 network allows us to train the entire multiscale model end-to-end and allows unfreezing the weights if a larger training set is available. We have experimented with unfreezing weights, but we quickly get overfitting problems with the available data material, this is therefore omitted from the paper.

Experiment 1 was conducted using ten training sets with a different number of triplets extracted from the same 220 WSI. From the result in Table 13.3, we see that the best performing model is trained on the largest dataset. However, the other models are not far behind. Even with a small value of N, the models do a good job at correctly predicting the WHO04 grade of WSIs.

Table 13.4: Tile-level prediction for each individual WSI in the segmentation test set, using the TRI_{WHO04}-5000-AUG model. The WSI numbering is referring to the WSIs in Fig. 13.6. The last row shows the average value and standard deviation for its respective column.

	Low-grade		High-grade	
WSI	Recall	F1-score	Recall	F1-Score
WSI A	-	-	0.79	0.88
WSI B	-	-	0.90	0.95
WSI C	-	-	0.86	0.92
WSI D	0.87	0.93	-	-
WSI E	-	-	0.94	0.97
WSI F	0.83	0.91	-	-
WSI G	0.86	0.92	-	-
WSI H	-	-	0.90	0.95
WSI I	0.85	0.92	-	-
WSI J	0.79	0.88	-	-
WSI K	-	-	0.92	0.96
WSI L	0.92	0.96	-	-
WSI M	0.68	0.81	-	-
WSI N	-	-	0.58	0.73
Average	0.83 ± 0.07	0.91 ± 0.04	0.84 ± 0.12	0.91 ± 0.08

Table 13.5: Aggregated tile-level result for all WSIs in the segmentation test set using the TRI_{WHO04}-5000-AUG model.

	Precision	Recall	F1-score
Low-grade	0.83	0.79	0.81
High-grade	0.90	0.81	0.85
Weighted Average	0.87	0.80	0.83

Table 13.6: Comparison table for automatic slide-level grading between our proposed method and the method presented in Jansen et al. [59]. Note that these results are based on models trained and evaluated on different datasets.

Model	Sensitivity	Specificity	Accuracy
$\text{TRI}_{\text{grade}}$	0.85	1.00	0.90
Jansen et al. [59]	0.71	0.76	0.74

Regarding overfitting, we tried training the models using unfrozen weights in the VGG16 networks, but this led to instantaneous overfitting of the model and had no improvements on the validation set. However, by freezing the weights, we see that all models improve on the validation dataset before reaching a plateau and eventually triggering the early stopping trigger. E.g., as shown in Fig. 13.7, the best model, $\text{TRI}_{\text{WHO04}}$-5000-AUG, improved its performance for seven epochs before training stopped after epoch 17. The weights from epoch seven were restored when using the model on the test sets. The number of trained epochs before the early stopping criteria is triggered decreases as the training dataset increases. This can be explained by the models trained on the larger datasets having more parameter updates per epoch than that of the smaller dataset models, thus reaching the plateau faster. Similarly, we see that the duration of one epoch is increasing as the dataset size increases. There is about a 60-hour difference in the smallest and largest model by comparing the total training time. Even though we would advise utilizing the most data to train a production model, it could be helpful to do an extended hyperparameter search and train multiple models on a smaller dataset.

Experiment 2, tile-level prediction, was conducted using the $\text{TRI}_{\text{WHO04}}$-5000-AUG model, which had a slide-level F1-score of 0.90. As seen in Fig. 13.6, Table 13.4 and 13.5, the results are overall excellent. The model does a very good job at correctly identifying both the low-grade and high-grade regions in the different WSIs. Table 13.4 shows that the model achieved an average F1-score of 91% for both the low-grade and high-grade classes. The aggregated score for all WSIs in Table 13.5 shows a small decrease in performance, with an F1-score of 81% and 85% for the two classes, respectively.

The largest misclassification in Fig. 13.6 is one of the regions in WSI-N, where the ground truth is high-grade, but the model predicts low-grade. When reevaluated by the pathologist, the misclassified area was

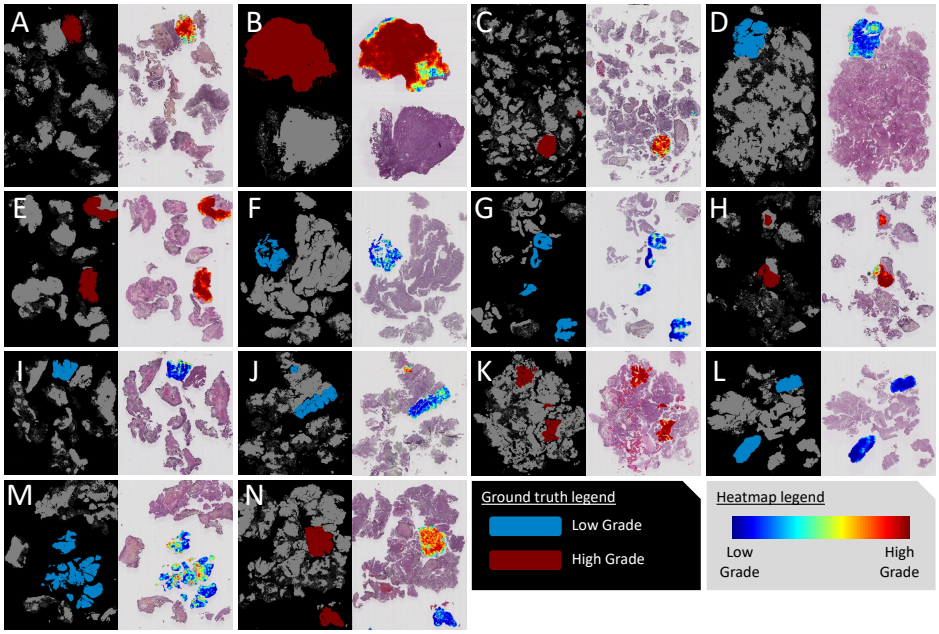


Figure 13.6: Ground truth annotations vs. model prediction. The WSI with a black background is the ground truth images with low- and high-grade annotations. The WSI with a grey background has superimposed a heatmap from the same area as the ground truth and highlights the predictions from the $\text{TRI}_{\text{WHO04}}$-model. For quantitative results, see Table 13.4 and 13.5.

found to be heterogenous, showing mixed low- and high-grade features, consequently regarded as high-grade initially. This illustrates one of the challenges with automatic grading of urothelial carcinoma, that grading between low- and high-grade is not two distinct binary classes but rather a continuous spectrum with a floating transition, making it difficult to set a hard threshold between the two.

To correct such misclassifications, and also avoid the costly task of annotating a large dataset, one possible solution is human-assisted learning. For example, the proposed $\text{TRI}_{\text{grade}}$ system could be used to find and predict urothelium regions into the low-grade and high-grade classes, e.g., like the regions seen in Fig. 13.6. Then, a pathologist could verify the regions in each WSI and correct misclassified regions. This way, a large, strongly labeled dataset could be created, and the $\text{TRI}_{\text{WHO04}}$-model could be fine-tuned on the new dataset.

A direct comparison of results with others reported in the literature is not

straightforward, as the experiments performed in this paper are conducted on a private dataset, which is often the case in many medical applications. To our knowledge, there exists no publically available NMIBC dataset or any publically available models from other researchers that we can evaluate on our dataset. The work of Jansen et al. [59] is based on the same labels but evaluated on a private dataset using different methods. Unfortunately, their models are not available for us to evaluate, and we do not have access to labels to train a Unet segmentation model from scratch, hence we cannot test the same approach by training the models ourselves. However, even though the dataset or model used in Jansen et al. [59] are not publically available, a comparison is still included as both research results are based on an NMIBC dataset of similar size (328 WSIs from 232 patients vs. our dataset of 300 WSIs), a similar split of the dataset into training, validation, and test, and the use of the same labels (WHO04). The results in Table 13.6 compare the slide-level sensitivity, specificity, and accuracy for our proposed $\text{TRI}_{\text{grade}}$ pipeline, to the results reported in table 3 from [59]. We achieve better results on all metrics, and with 45 of the 50 WSIs correctly predicted, we achieve an accuracy of 0.90.

Training and validation accuracy from the training of the $\text{TRI}_{\text{WHO04}}$-5000-AUG model is shown in Fig. 13.7. The model uses frozen pre-trained weights for the VGG16 networks, and only the last layers in the model have random weights which are being optimized. The model uses the largest training dataset from Table 13.2 with a mini-batch size of 128, resulting in a large number of weight updates per epoch, and the majority of the accuracy is achieved from the first epoch. After the initial epoch, the validation accuracy is not improving too much. This is most likely because the datasets use imprecise weak labels (e.g., all urothelium triplets extracted from a high-grade WSI will have the class label high-grade, but not all triplets from this WSI will represent high-grade tissue). Note also that all the urothelium triplets from all the WSIs in the validation set are predicted before Tensorflow computes the accuracy score for the validation set.

13.5.1 Usage scenarios

The automatic $\text{TRI}_{\text{grade}}$ system presented in this paper has many potential applications. The tissue model we presented in Wetteland et al. [140] provides the tissue segmentation maps, which clinicians can use to discriminate urothelium regions from other tissue classes. This can be a valuable tool

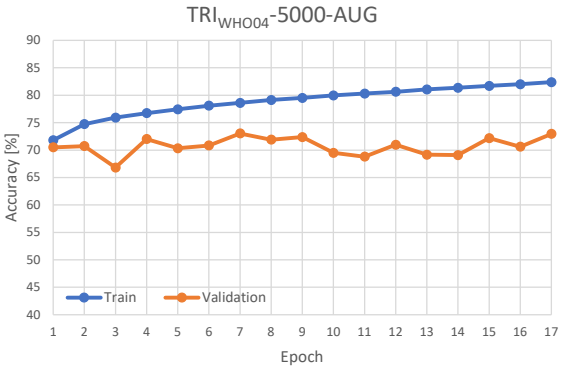


Figure 13.7: Training and validation accuracy for the TRI$_{WHO04}$-5000-AUG model. The model is trained on imprecise weak labels, using the largest training set in Table 13.2. Results are shown for tile-level prediction on the entire training and validation sets. Validation accuracy is computed at the end of each epoch.

to aid pathologists in examining the whole-slide images by focusing their attention on the diagnostic relevant areas of the stained specimen. With the addition of the TRI$_{WHO04}$-model presented in this paper, the focus can not only be aimed towards the urothelium regions in general but be further narrowed down to the most *severe* urothelium regions.

The automated slide-level prediction can potentially be used to prioritize high-grade patients for earlier examination. Also, it can be used as input to an automatic prognostic tool and output a measure of the patient's overall clinical outcome, such as the risk of recurrence, 1-yr and 5-yr survival rate, and mortality. In the future, it is also a possibility to use it in an automatic system that predicts how a patient will respond to a given treatment and therapy program.

13.5.2 Limitations

In the paper, we train a model to classify urothelium tissue into two classes, low- and high-grade carcinoma. However, it is also a possibility that the urothelium tissue can be healthy non-cancerous tissue. Since our models are dependent on the weak slide-level label, and all cases in the data material are diagnosed with cancer, we currently do not have any training material containing non-cancerous samples.

All WSIs in this study are collected from the same laboratory and consists of high quality with relatively small variations in stain colors and little blur.

This is both a strength in the sense that we have produced good models and reliable predictions, but also a limitation in the sense that we do not know how the system will perform on slides of lower quality.

13.5.3 Future work

In future work, preprocessing steps might be added to deal with color variations, blur, and folded tissue, or the tissue segmentation model can be updated with a new class for blur, providing a more generalized system.

From [140] it was concluded that for the tissue segmentation task, the multiscale TRI-25x-100x-400x model (which is used as the $\text{TRI}_{\text{tissue}}$-model in this work) provided the best performance. Following, a multiscale model was adopted for the grading task as well, with the masking of the urothelium tissue performed at the 400x level. However, the large field-of-view provided by the 25x and 100x magnification will bring neighboring tissue types into the triplet, like, for example, damaged tissue, which might affect the performance in such areas. In future work, we would like to use the tissue segmentation maps and not only extract the urothelium tissue but also mask out unwanted regions of damaged tissue and blood. Incorporating attention modules is also something we will try, which would further help explain what parts of the WSI are responsible for the predictions.

Cells of low-grade cancer often resemble that of non-cancerous cells, and high-grade cells have a more abnormal appearance and are disorganized. Thus, we expect that non-cancerous tissue would be predicted as low-grade carcinoma. However, this is our expectation as we do not have verified material to test this on. To better detect these non-cancerous regions in the future, we would have to expand our training dataset to include examples of non-cancerous urothelium. The $\text{TRI}_{\text{WHO04}}$-model architecture must be updated to include one additional class on the output and then be trained on the updated dataset.

The proposed model uses three VGG16 networks as feature extractors. In the future, we would like to experiment with other deep learning networks for our base model. Newer deep learning models continuously improve the results on datasets like ImageNet, and could potentially improve feature extraction of urothelium tissue. We also plan to look into different ways of fusing the multiscale information, both for the tissue classifier ($\text{TRI}_{\text{tissue}}$) and grade-classifier ($\text{TRI}_{\text{WHO04}}$).

13.6 Conclusion

In this paper, we have proposed a $\text{TRI}_{\text{grade}}$ pipeline for automatic grading of urothelial carcinoma slides based on the WHO04 grading system. First, the slide is segmented into the tissue classes (urothelium, stroma, muscle, blood, damaged tissue, and background). Next, tiles are extracted at three magnification levels (25x, 100x, and 400x) from the urothelium regions. The three tiles form a triplet, which is fed sequentially to a multiscale CNN-based WHO04 grading model.

The proposed method will generate a tissue segmentation map, helpful for the clinicians to easier find diagnostic relevant regions during an examination. The system will also output a WHO04 grade heatmap, highlighting the most severe urothelium tissue regions, beneficial for the pathologists who can focus their limited per-patient time on the most important regions in the WSI. Finally, the system produces a slide-level WHO04 grade that could potentially be used to prioritize high-grade patients for earlier examination, as well as suggest the diagnosis to the pathologist.

Ten WHO04 grade models were trained on datasets of varying sizes. Note that all the same number of WSI were used all the time, but a different number of triplets were extracted from each WSI, constituting the training set. The model trained on the largest training dataset achieved the best result, a weighted average F1-score of 0.90 on the test set. This model was further evaluated on a segmentation test set, where low- and high-grade regions were annotated by a pathologist. On this task, the model got an average F1-score of 0.91 on both the low-grade and high-grade classes.

The system as a whole can be used by clinicians and pathologists to potentially improve their decision-making and further help patients by receiving correct diagnoses and treatment.

Bibliography

[1] **Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, et al.** Tensorflow: A system for large-scale machine learning. In: *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)* 265–283. 2016.

[2] **Amann J, Blasimme A, Vayena E, Frey D, Madai VI**. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making* 20: 1–9. 2020.

[3] **Anastasiadis A, de Reijke TM**. Best practice in the treatment of nonmuscle invasive bladder cancer. *Therapeutic advances in urology* 4: 13–32. 2012.

[4] **Antoni S, Ferlay J, Soerjomataram I, Znaor A, Jemal A, Bray F**. Bladder cancer incidence and mortality: a global overview and recent trends. *European urology* 71: 96–108. 2017.

[5] **Araújo T, Aresta G, Castro E, Rouco J, Aguiar P, Eloy C, Polónia A, Campilho A**. Classification of breast cancer histology images using convolutional neural networks. *PloS one* 12: e0177544. 2017.

[6] **Babjuk M**. Transurethral resection of non–muscle-invasive bladder cancer. *European Urology Supplements* 8: 542–548. 2009.

[7] **Babjuk M, Böhle A, Burger M, Capoun O, Cohen D, Compérat EM, Hernández V, Kaasinen E, Palou J, Rouprêt M, et al.** Eau guidelines on non–muscle-invasive urothelial carcinoma of the bladder: update 2016. *European urology* 71: 447–461. 2017.

[8] **Baldi P**. Autoencoders, unsupervised learning, and deep architectures. In: *Proceedings of ICML workshop on unsupervised and transfer learning* 37–49. 2012.

[9] **Benjamens S, Dhunnoo P, Meskó B**. The state of artificial intelligence-based fda-approved medical devices and algorithms: an online database. *NPJ digital medicine* 3: 1–8. 2020.

[10] **Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A**. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. *Nature reviews Clinical oncology* 16: 703–715. 2019.

[11] **BigTIFF**. Extending libtiff library with support for the new BigTIFF format. *Available at* `http://simplesystems.org/libtiff/bigtiffpr.html`. Last accessed 07.09.2021. 2021.

[12] **Botteman MF, Pashos CL, Redaelli A, Laskin B, Hauser R**. The health economics of bladder cancer. *Pharmacoeconomics* 21: 1315–1330. 2003.

[13] **Boulila W, Sellami M, Driss M, Al-Sarem M, Safaei M, Ghaleb FA**. RS-DCNN: A novel distributed convolutional-neural-networks based-approach for big remote-sensing image classification. *Computers and Electronics in Agriculture* 182: 106014. 2021.

[14] **Brancati N, De Pietro G, Riccio D, Frucci M**. Gigapixel histopathological image analysis using attention-based neural networks. *arXiv preprint arXiv:210109992*. 2021.

[15] **Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A**. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* 68: 394–424. 2018.

[16] **Browning L, Fryer E, Roskell D, White K, Colling R, Rittscher J, Verrill C**. Role of digital pathology in diagnostic histopathology in the response to covid-19: results from a survey of experience in a uk tertiary referral hospital. *Journal of clinical pathology* 74: 129–132. 2021.

[17] **Buetti-Dinh A, Galli V, Bellenberg S, Ilie O, Herold M, Christel S, Boretska M, Pivkin IV, Wilmes P, Sand W, et al.** Deep neural networks outperform human expert's capacity in characterizing bioleaching bacterial biofilm composition. *Biotechnology Reports* 22: e00321. 2019.

[18] **C Tao, H Pan, Y Li, Z Zou**. Unsupervised spectral–spatial feature learning with stacked sparse autoencoder for hyperspectral imagery classification. *IEEE Geoscience and Remote Sensing Letters* 12: 2438–2442. 2015.

[19] **Campanella G, Hanna MG, Geneslaw L, Miraflor A, Silva VWK, Busam KJ, Brogi E, Reuter VE, Klimstra DS, Fuchs TJ**. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine* 25: 1301–1309. 2019.

[20] **Cancer Registry of Norway**. Cancer in norway 2005. *Cancer incidence, mortality, survival and prevalence in Norway* 18. 2006.

[21] **Cancer Registry of Norway**. Cancer in norway 2010. *Cancer incidence, mortality, survival and prevalence in Norway* 26. 2012.

[22] **Cancer Registry of Norway**. Cancer in norway 2015. *Cancer incidence, mortality, survival and prevalence in Norway* 28. 2016.

[23] **Cancer Registry of Norway**. Cancer in norway 2018. *Cancer incidence, mortality, survival and prevalence in Norway* 20. 2019.

[24] **Ceccopieri C, Skonieczna J, Madej JP**. Modification of a haematoxylin, eosin, and natural saffron staining method for the detection of connective tissue. *Journal of Veterinary Research* 65: 125. 2021.

[25] **Celik Y, Talo M, Yildirim O, Karabatak M, Acharya UR**. Automated invasive ductal carcinoma detection based using deep transfer learning with whole-slide images. *Pattern Recognition Letters* 133: 232–239. 2020.

[26] **Chan JK**. The wonderful colors of the hematoxylin–eosin stain in diagnostic surgical pathology. *International journal of surgical pathology* 22: 12–32. 2014.

[27] **Cheplygina V, de Bruijne M, Pluim JP**. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis* 54: 280–296. 2019.

[28] **Chini M, Hostache R, Giustarini L, Matgen P**. A hierarchical split-based approach for parametric thresholding of sar images: Flood inundation as a test case. *IEEE Transactions on Geoscience and Remote Sensing* 55: 6975–6988. 2017.

[29] **Chollet F, et al.** Keras. https://github.com/fchollet/keras. Last accessed 06.08.2021. 2015.

[30] **Cui J, Gong K, Guo N, Wu C, Meng X, Kim K, Zheng K, Wu Z, Fu L, Xu B, et al.** Pet image denoising using unsupervised deep learning. *European journal of nuclear medicine and molecular imaging* 46: 2780–2789. 2019.

[31] **Dalheim ON**. *Semi-Supervised Image Segmentation of Medical Data*. Master's thesis University of Stavanger, Norway. 2020.

[32] **Dalheim ON, Wetteland R, Kvikstad V, Janssen EAM, Engan K**. Semi-supervised tissue segmentation of histological images. *Colour and Visual Computing Symposium* 2688. 2020.

[33] **Daniel N, Larey A, Aknin E, Osswald GA, Caldwell JM, Rochman M, Collins MH, Yang GY, Arva NC, Capocelli KE, et al.** Pecnet: A deep multi-label segmentation network for eosinophilic esophagitis biopsy diagnostics. *arXiv preprint arXiv:210302015*. 2021.

[34] **Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L**. Imagenet: A large-scale hierarchical image database. In: *2009 IEEE conference on computer vision and pattern recognition* 248–255. Ieee. 2009.

[35] **Doyle S, Feldman M, Tomaszewski J, Madabhushi A**. A boosted bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies. *IEEE transactions on biomedical engineering* 59: 1205–1218. 2010.

[36] **E Edston, L Gröntoft**. Saffron—a connective tissue counterstain in routine pathology. *Journal of Histotechnology* 20: 123–125. 1997.

[37] **Eble JN, Sauter G, Epstein JI, Sesterhenn IA**. World Health Organization Classification of Tumours. Pathology and Genetics of Tumours of the Urinary System and Male Genital Organs. *IARC Press: Lyon*. 2004.

[38] **Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I, Bray F**. Global cancer observatory: Cancer today. *Available at https://gco.iarc.fr/today*. Last accessed 01.10.2021. 2020.

[39] **Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM**. Estimates of worldwide burden of cancer in 2008: Globocan 2008. *International journal of cancer* 127: 2893–2917. 2010.

[40] **Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F**. Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. *International journal of cancer* 136: E359–E386. 2015.

[41] **Forman G, Scholz M**. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *ACM SIGKDD Explorations Newsletter* 12: 49–57. 2010.

[42] **G Litjens, T Kooi, B E Bejnordi et al**. A survey on deep learning in medical image analysis. *Medical Image Analysis* 42: 60–88. 2017.

[43] **Goldstein M, Uchida S**. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS one* 11: e0152173. 2016.

[44] **Goodfellow I, Bengio Y, Courville A**. *Deep learning*. London: MIT press. 2016.

[45] **Graham S, Vu QD, Raza SEA, Azam A, Tsang YW, Kwak JT, Rajpoot N**. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis* 58: 101563. 2019.

[46] **Graham S, Vu QD, Raza SEA, Azam A, Tsang YW, Kwak JT, Rajpoot N**. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis* 58: 101563. 2019.

[47] **Guo Z, Liu H, Ni H, Wang X, Su M, Guo W, Wang K, Jiang T, Qian Y**. A fast and refined cancer regions segmentation framework in whole-slide breast pathological images. *Scientific reports* 9: 1–10. 2019.

[48] **Gupta A, Duggal R, Gehlot S, Gupta R, Mangal A, Kumar L, Thakkar N, Satpathy D**. Gcti-sn: Geometry-inspired chemical and tissue invariant stain normalization of microscopic medical images. *Medical Image Analysis* 65: 101788. 2020.

[49] **Halicek M, Shahedi M, Little JV, Chen AY, Myers LL, Sumer BD, Fei B**. Head and neck cancer detection in digitized whole-slide histology using convolutional neural networks. *Scientific reports* 9: 1–11. 2019.

[50] **Hamamatsu**. NDPI - image format for microscopes. *Available at https://fileinfo.com/extension/ndpi*. Last accessed 07.09.2021. 2021.

[51] **Hanna MG, Reuter VE, Ardon O, Kim D, Sirintrapun SJ, Schüffler PJ, Busam KJ, Sauter JL, Brogi E, Tan LK, et al.** Validation of a digital pathology system including remote review during the covid-19 pandemic. *Modern Pathology* 33: 2115–2127. 2020.

[52] **Hashimoto N, Fukushima D, Koga R, Takagi Y, Ko K, Kohno K, Nakaguro M, Nakamura S, Hontani H, Takeuchi I**. Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 3852–3861. 2020.

[53] **He K, Zhang X, Ren S, Sun J**. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: *Proceedings of the IEEE international conference on computer vision* 1026–1034. 2015.

[54] **Hosseinimotlagh S, Papalexakis EE**. Unsupervised content-based identification of fake news articles with tensor decomposition ensembles. In: *Proceedings of the Workshop on Misinformation and Misbehavior Mining on the Web (MIS2)*. 2018.

[55] **Hou L, Samaras D, Kurc TM, Gao Y, Davis JE, Saltz JH**. Patch-based convolutional neural network for whole slide tissue image classification. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* 2424–2433. 2016.

[56] **Huang X, He H, Wei P, Zhang C, Zhang J, Chen J**. Tumor tissue segmentation for histopathological images. *Proceedings of the ACM Multimedia Asia* 1–4. 2019.

[57] **Ianni JD, Soans RE, Sankarapandian S, Chamarthi RV, Ayyagari D, Olsen TG, Bonham MJ, Stavish CC, Motaparthi K, Cockerell CJ, et al.** Tailored for real-world: a whole slide image classification system validated on uncurated multi-site data emulating the prospective pathology workload. *Scientific reports* 10: 1–12. 2020.

[58] **Janowczyk A, Zuo R, Gilmore H, Feldman M, Madabhushi A**. Histoqc: an open-source quality control tool for digital pathology slides. *JCO clinical cancer informatics* 3: 1–7. 2019.

[59] **Jansen I, Lucas M, Bosschieter J, de Boer OJ, Meijer SL, van Leeuwen TG, Marquering HA, Nieuwenhuijzen JA, de Bruin DM, Savci-Heijink CD**. Automated detection and grading of non–muscle-invasive urothelial cell carcinoma of the bladder. *The American journal of pathology* 190: 1483–1490. 2020.

[60] **Javed S, Mahmood A, Werghi N, Benes K, Rajpoot N**. Multiplex cellular communities in multi-gigapixel colorectal cancer histology images for tissue phenotyping. *IEEE Transactions on Image Processing* 29: 9204–9219. 2020.

[61] **K Dercksen, W Bulten, G Litjens**. Dealing with label scarcity in computational pathology: A use case in prostate cancer classification. *Proceedings of Machine Learning Research – Accepted :1–4, 2019, Extended Abstract – MIDL 2019 submission.* 2019.

[62] **Kather JN, Weis CA, Bianconi F, Melchers SM, Schad LR, Gaiser T, Marx A, Zöllner FG**. Multi-class texture analysis in colorectal cancer histology. *Scientific reports* 6: 27988. 2016.

[63] **Kohlberger T, Liu Y, Moran M, Chen PHC, Brown T, Hipp JD, Mermel CH, Stumpe MC**. Whole-slide image focus quality: Automatic assessment and impact on ai cancer detection. *Journal of pathology informatics* 10. 2019.

[64] BLÆREKREFT. *Available at* `https://www.kreftregisteret.no/Temasider/kreftformer/blarekreft/`. Last accessed 29.04.2020.

[65] **Kvikstad V, Mangrud OM, Gudlaugsson E, Dalen I, Espeland H, Baak JP, Janssen EAM**. 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.

[66] **LeCun Y, Bengio Y, Hinton G**. Deep learning. *nature* 521: 436. 2015.

[67] **Leica Biosystem**. Aperio ImageScope - pathology slide viewing software. *Available at* `https://www.leicabiosystems.com/digital-pathology/manage/aperio-imagescope/`. Last accessed 07.09.2021. 2021.

[68] **Li J, Sarma KV, Ho KC, Gertych A, Knudsen BS, Arnold CW**. A multi-scale u-net for semantic segmentation of histological images from radical prostatectomies. In: *AMIA Annual Symposium Proceedings* vol. 2017 1140. American Medical Informatics Association. 2017.

[69] **Li R, Huang J**. Fast regions-of-interest detection in whole slide histopathology images. In: *International Workshop on Patch-based Techniques in Medical Imaging* 120–127. Springer. 2015.

[70] **Li X, Chen H, Qi X, Dou Q, Fu CW, Heng PA**. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE transactions on medical imaging* 37: 2663–2674. 2018.

[71] **Li Z, Tao R, Wu Q, Li B**. Da-refinenet: A dual input wsi image segmentation algorithm based on attention. *arXiv preprint arXiv:190706358*. 2019.

[72] **Ligthart A, Catal C, Tekinerdogan B**. Analyzing the effectiveness of semi-supervised learning approaches for opinion spam classification. *Applied Soft Computing* 101: 107023. 2021.

[73] **Liu S, Ren J, Chen Z, Hu K, Xiao F, Li X, Gao X**. EffiDiag: an efficient framework for breast cancer diagnosis in multi-gigapixel whole slide images. In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 663–669. IEEE. 2020.

[74] **Lo YC, Juang CF, Chung IF, Guo SN, Huang ML, Wen MC, Lin CJ, Lin HY**. Glomerulus detection on light microscopic images of renal pathology with the faster r-cnn. In: *International Conference on Neural Information Processing* 369–377. Springer. 2018.

[75] **Long J, Shelhamer E, Darrell T**. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* 3431–3440. 2015.

[76] **Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, Abraham J, Adair T, Aggarwal R, Ahn SY, et al.** Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the global burden of disease study 2010. *The lancet* 380: 2095–2128. 2012.

[77] **Lu MY, Chen RJ, Wang J, Dillon D, Mahmood F**. Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding. *arXiv preprint arXiv:191010825*. 2019.

[78] **Lucas M, Jansen I, van Leeuwen TG, Oddens JR, de Bruin DM, Marquering HA**. Deep learning–based recurrence prediction in patients with non–muscle-invasive bladder cancer. *European Urology Focus*. 2020.

[79] **Madabhushi A, Lee G**. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical image analysis* 33: 170–175. 2016.

[80] **Malkenes Ø**. *Image processing on histopathological images of urothelial carcinoma–assessment of immune cells*. Master's thesis University of Stavanger, Norway. 2018.

[81] **Mangrud OM**. *Identification of patients with high and low risk of progresson of urothelial carcinoma of the urinary bladder stage Ta and T1*. Ph.D. thesis PhD thesis, Ph. D. dissertation, University of Bergen. 2014.

[82] **Mangrud OM, Waalen R, Gudlaugsson E, Dalen I, Tasdemir I, Janssen EAM, Baak JP**. Reproducibility and prognostic value of who1973 and who2004 grading systems in tat1 urothelial carcinoma of the urinary bladder. *PLoS One* 9: e83192. 2014.

[83] **Märkl B, Füzesi L, Huss R, Bauer S, Schaller T**. Number of pathologists in germany: Comparison with european countries, usa, and canada. *Virchows Archiv* 478: 335–341. 2021.

[84] **Martinez K, Cupitt J**. Vips-a highly tuned image processing software architecture. In: *IEEE International Conference on Image Processing 2005* vol. 2 II–574. IEEE. 2005.

[85] **Medical Imaging and Technology Alliance**. DICOM - digital imaging and communications in medicine. *Available at* `https://www.dicomstandard.org/`. Last accessed 2021-07-09. 2021.

[86] **Morales S, Engan K, Naranjo V**. Artificial intelligence in computational pathology–challenges and future directions. *Digital Signal Processing* 103196. 2021.

[87] **Muhammad H, Sigel CS, Campanella G, Boerner T, Pak LM, Büttner S, IJzermans JN, Koerkamp BG, Doukas M, Jarnagin WR, et al.** Towards unsupervised cancer subtyping: predicting prognosis using a histologic visual dictionary. *arXiv preprint arXiv:190305257*. 2019.

[88] **Munir K, Elahi H, Ayub A, Frezza F, Rizzi A**. Cancer diagnosis using deep learning: a bibliographic review. *Cancers* 11: 1235. 2019.

[89] **Nesse AB**. *Classifying Dinoflagellates in Palynological Slides Using Convolutional Neural Networks*. Master's thesis University of Stavanger, Norway. 2020.

[90] **Niazi MKK, Parwani AV, Gurcan MN**. Digital pathology and artificial intelligence. *The lancet oncology* 20: e253–e261. 2019.

[91] **Norwegian Artificial Intelligence Research Consortium (NORA)**. MedAI: Transparency in Medical Image Segmentation. *Available at https://www.nora.ai/Competition/image-segmentation.html*. Last accessed 06.08.2021. 2021.

[92] **Olivas ES, Guerrero JDM, Martinez-Sober M, Magdalena-Benedito JR, Serrano L, et al.** *Handbook of research on machine learning applications and trends: Algorithms, methods, and techniques: Algorithms, methods, and techniques*. Pennsylvania: IGI Global. 2009.

[93] **Ouali Y, Hudelot C, Tami M**. An overview of deep semi-supervised learning. *arXiv preprint arXiv:200605278*. 2020.

[94] **Pan SJ, Yang Q**. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22: 1345–1359. 2010.

[95] **Parkin DM, Bray F, Ferlay J, Pisani P**. Estimating the world cancer burden: Globocan 2000. *International journal of cancer* 94: 153–156. 2001.

[96] **Parkin DM, Bray F, Ferlay J, Pisani P**. Global cancer statistics, 2002. *CA: a cancer journal for clinicians* 55: 74–108. 2005.

[97] **Parkin DM, Pisani P, Ferlay J**. Global cancer statistics. *CA: a cancer journal for clinicians* 49: 33–64. 1999.

[98] **Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al.** Scikit-learn: Machine learning in python. *Journal of machine learning research* 12: 2825–2830. 2011.

[99] **Peikari M, Salama S, Nofech-Mozes S, Martel AL**. A cluster-then-label semi-supervised learning approach for pathology image classification. *Scientific reports* 8: 1–13. 2018.

[100] **Poloprutskỳ Z, Soukup P**. Analytical maps as a basis for understanding the development of rural architecture. *AUC Geographica* 56: 31–43. 2021.

[101] **PyVips**. Binding for the libvips image processing library. *Available at* `https://pypi.org/project/pyvips/`. Last accessed 07.09.2021. 2021.

[102] **PyVips**. Speed and memory benchmark. *Available at* `https://github.com/libvips/libvips/wiki/Speed-and-memory-use`. Last accessed 07.09.2021. 2021.

[103] **Refaeilzadeh P, Tang L, Liu H**. Cross-validation. *Encyclopedia of database systems* 5: 532–538. 2009.

[104] **Ronneberger O, Fischer P, Brox T**. U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical image computing and computer-assisted intervention* 234–241. Springer. 2015.

[105] **Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, et al.** Imagenet large scale visual recognition challenge. *International journal of computer vision* 115: 211–252. 2015.

[106] **Saginala K, Barsouk A, Aluru JS, Rawla P, Padala SA, Barsouk A**. Epidemiology of bladder cancer. *Medical Sciences* 8: 15. 2020.

[107] **Salehi P, Chalechale A**. Pix2pix-based stain-to-stain translation: A solution for robust stain normalization in histopathology images analysis. In: *2020 International Conference on Machine Vision and Image Processing (MVIP)* 1–7. IEEE. 2020.

[108] **Sellaro TL, Filkins R, Hoffman C, Fine JL, Ho J, Parwani AV, Pantanowitz L, Montalto M**. Relationship between magnification and resolution in digital pathology systems. *Journal of pathology informatics* 4. 2013.

[109] **Senaras C, Niazi MKK, Lozanski G, Gurcan MN**. Deepfocus: detection of out-of-focus regions in whole slide digital images using deep learning. *PloS one* 13: e0205387. 2018.

[110] **Shutov A, Maleev A, Zhuravlev V**. Complex quasiperiodic self-similar tilings: their parameterization, boundaries, complexity, growth and symmetry. *Acta Crystallographica Section A: Foundations of Crystallography* 66: 427–437. 2010.

[111] **Sievert KD, Amend B, Nagele U, Schilling D, Bedke J, Horstmann M, Hennenlotter J, Kruck S, Stenzl A**. Economic aspects of bladder cancer: what are the benefits and costs? *World journal of urology* 27: 295–300. 2009.

[112] **Silva-Rodriguez J, Colomer A, Dolz J, Naranjo V**. Self-learning for weakly supervised gleason grading of local patterns. *IEEE journal of biomedical and health informatics*. 2021.

[113] **Simonyan K, Zisserman A**. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:14091556*. 2014.

[114] **Sirinukunwattana K, Alham NK, Verrill C, Rittscher J**. Improving whole slide segmentation through visual context-a systematic study. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* 192–200. Springer. 2018.

[115] **Skrede OJ, De Raedt S, Kleppe A, Hveem TS, Liestøl K, Maddison J, Askautrud HA, Pradhan M, Nesheim JA, Albregtsen F, et al.** Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *The Lancet* 395: 350–360. 2020.

[116] **Stavanger Aftenblad**. Pasienter må vente åtte uker på prøvesvar. *Available at https://www.aftenbladet.no/lokalt/i/Wk332/pasienter-ma-vente-atte-uker-pa-prvesvar*. 2020.

[117] **Ström P, Kartasalo K, Olsson H, Solorzano L, Delahunt B, Berney DM, Bostwick DG, Evans AJ, Grignon DJ, Humphrey PA, et al.** Pathologist-level grading of prostate biopsies with artificial intelligence. *arXiv preprint arXiv:190701368*. 2019.

[118] **Sulimowicz L, Ahmad I**. "rapid" regions-of-interest detection in big histopathological images. In: *Multimedia and Expo (ICME), 2017 IEEE International Conference on* 595–600. IEEE. 2017.

[119] **Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F**. 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.

[120] **Svendsen F**. *Image Processing and Deep Neural Networks for Detection of Immune Cells on Histological Images of Bladder Cancer*. Master's thesis University of Stavanger, Norway. 2019.

[121] **Tellez D, Litjens G, Bándi P, Bulten W, Bokhorst JM, Ciompi F, van der Laak J**. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical image analysis* 58: 101544. 2019.

[122] **Tharwat A**. Classification assessment methods. *Applied Computing and Informatics*. 2020.

[123] Bladder, source: Globocan 2018. *Available at https://gco.iarc.fr/today/data/factsheets/cancers/30-Bladder-fact-sheet.pdf*. Last accessed 15.04.2020.

[124] **The Medical Futurist**. FDA-approved A.I.-based algorithms. *Available at https://medicalfuturist.com/fda-approved-ai-based-algorithms/*. Last accessed 06.08.2021. 2021.

[125] **UK CR**. Diagram showing the t stages of bladder cancer. *Available at https://commons.wikimedia.org/wiki/File:Diagram_showing_the_T_stages_of_bladder_cancer_CRUK_372.svg*. Last accessed 30.09.2021. 2016.

[126] **Urdal J**. *Image processing and classification of urothelial carcinoma using tissue sample images*. Master's thesis University of Stavanger, Norway. 2016.

[127] **Urdal J, Engan K, Kvikstad V, Janssen EAM**. Prognostic prediction of histopathological images by local binary patterns and rusboost. In: *2017 25th European Signal Processing Conference (EUSIPCO)* 2349–2353. IEEE. 2017.

[128] **Urteaga J**. *Regional Convolutional Neural Network for Cell Detection and Classification in Urinary Bladder Cancer*. Master's thesis University of Stavanger, Norway. 2020.

[129] **V Cheplygina, M de Bruijne, J PW Pluim**. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis* 280–297. 2019.

[130] **Villamizar M, Garrell A, Sanfeliu A, Moreno-Noguer F**. Online human-assisted learning using random ferns. In: *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)* 2821–2824. IEEE. 2012.

[131] **Vu QD, Graham S, Kurc T, To MNN, Shaban M, Qaiser T, Koohbanani NA, Khurram SA, Kalpathy-Cramer J, Zhao T, et al.** Methods for segmentation and classification of digital microscopy tissue images. *Frontiers in bioengineering and biotechnology* 7: 53. 2019.

[132] **Wagner SJ, Khalili N, Sharma R, Boxberg M, Marr C, Back Wd, Peng T**. Structure-preserving multi-domain stain color augmentation using style-transfer with disentangled representations. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* 257–266. Springer. 2021.

[133] **Wang D, Khosla A, Gargeya R, Irshad H, Beck AH**. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:160605718*. 2016.

[134] **Wang J, MacKenzie JD, Ramachandran R, Chen DZ**. A deep learning approach for semantic segmentation in histology tissue images. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* 176–184. Springer. 2016.

[135] **Wang L**. Discovering phase transitions with unsupervised learning. *Physical Review B* 94: 195105. 2016.

[136] **Wetteland R**. *Classification of histological images of bladder cancer using deep learning*. Master's thesis University of Stavanger, Norway. 2017.

[137] **Wetteland R, Engan K, Eftestøl T**. Parameterized extraction of tiles in multilevel gigapixel images. In: *Accepted for publication, 12th International Symposium on Image and Signal Processing and Analysis (ISPA 2021)*. IEEE. 2021.

[138] **Wetteland R, Engan K, Eftestøl T, Kvikstad V, Janssen EAM**. Multiclass tissue classification of whole-slide histological images using convolutional neural networks. In: *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods - Volume 1: ICPRAM* 320–327. INSTICC Prague: SciTePress. 2019.

[139] **Wetteland R, Engan K, Eftestøl T, Kvikstad V, Janssen EAM**. Multiscale Deep Neural Networks for Multiclass Tissue Classification of Histological Whole-Slide Images. *Medical Imaging with Deep Learning: MIDL 2019 – Extended Abstract Track* arXiv:1909.01178. 2019.

[140] **Wetteland R, Engan K, Eftestøl T, Kvikstad V, Janssen EAM**. A multiscale approach for whole-slide image segmentation of five tissue classes in urothelial carcinoma slides. *Technology in Cancer Research and Treatment (TCRT)* 19. 2020.

[141] **Wu S, Rupprecht C, Vedaldi A**. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 1–10. 2020.

[142] **Xu H, Park S, Lee SH, Hwang TH**. Using transfer learning on whole slide images to predict tumor mutational burden in bladder cancer patients. *bioRxiv* 554527. 2019.

[143] **Y Song, C Zhang, J Lee et al**. Semi-supervised discriminative classification with application to tumorous tissues segmentation of mr brain images. *Pattern Analysis and Applications* 12: 99–115. 2009.

[144] **Yadav S, Shukla S**. Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In: *2016 IEEE 6th International conference on advanced computing (IACC)* 78–83. IEEE. 2016.

[145] **Yao J, Boben M, Fidler S, Urtasun R**. Real-time coarse-to-fine topologically preserving segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2947–2955. 2015.

[146] **Yari Y, Nguyen H**. A state-of-the-art deep transfer learning-based model for accurate breast cancer recognition in histology images. In: *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)* 900–905. IEEE. 2020.

[147] **Z Shi, L He, K Suzuki, T Nakamura, H Itoh**. Survey on neural networks used for medical image processing. *International journal of computational science* 3: 86–100. 2009.

[148] **Zeiss**. CZI - image format for microscopes. *Available at* `https://www.zeiss.com/microscopy/int/products/microscope-software/zen/czi.html`. Last accessed 07.09.2021. 2021.

[149] **Zhang H, Cissé M, Dauphin YN, Lopez-Paz D**. mixup: Beyond empirical risk minimization. *CoRR* abs/1710.09412. 2017.

[150] **Zhang K, Crookes D, Diamond J, Fei M, Wu J, Zhang P, Zhou H**. Multi-scale colorectal tumour segmentation using a novel coarse to fine strategy. In: *BMVC*. 2016.

[151] **Zhang Z, Chen P, McGough M, Xing F, Wang C, Bui M, Xie Y, Sapkota M, Cui L, Dhillon J, et al.** Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nature Machine Intelligence* 1: 236–245. 2019.

[152] **Zhou L, Wei W**. Dic: deep image clustering for unsupervised image segmentation. *IEEE Access* 8: 34481–34491. 2020.