UNIVERSITY OF CALIFORNIA

Los Angeles

Large-scale Whole Slide Image Analysis with Deep Learning Methods to Improve Prostate Cancer Diagnosis

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Bioengineering

by

Jiayun Li

2021

ABSTRACT OF THE DISSERTATION

Large-scale Whole Slide Image Analysis with Deep Learning Methods to Improve Prostate Cancer Diagnosis

by

Jiayun Li
Doctor of Philosophy in Bioengineering
University of California, Los Angeles, 2021
Professor Denise R. Aberle, Co-Chair
Professor Corey W. Arnold, Co-Chair

Gleason grading system serves as an essential component in risk stratification and treatment planning for prostate cancer patients. Currently, Gleason grading relies on pathologists to examine glass tissue slides at scanning magnification and localize suspicious regions for higher power examination. Such process can be time-consuming and prone to inter- and intra- observer variability. Moreover, the Gleason grading system may be constrained by its categorization system, which cannot fully capture the disease's continuous feature spectrum. With the recent development of digital slide scanners and the approval of using digitized slides for primary diagnosis by the Food and Drug Administrations (FDA), large numbers of traditional glass slides have been digitized into high resolution whole slide images (WSIs), opening opportunities in developing computational image analysis tools to reduce pathologists' workload and potentially improve inter- and intra- observer agreement.

This research attempts to address several challenges in WSI analysis and assist the histopathologic evaluation of prostate cancer. A tile-level semantic segmentation model, which explicitly leverages multi-scale representations, is first proposed to generate pixel-wise Gleason pattern predictions and facilitate estimation of percentage of different patterns. An expectation-maximization (EM)-based semi-supervised learning framework is then developed to exploit information embedded in weakly-labeled samples to further improve the performance of segmentation models, which alleviates the need of expensive pixel-wise annotations. Besides building segmentation tools for tiles extracted from WSIs, a novel multi-resolution multiple instance learning-based model, which can be trained with slide-level labels, is proposed to identify informative regions and provide slide-level Gleason grade group prediction. The model is developed and validated on a large-scale prostate biopsy dataset. Furthermore, a deep learning system, which leverages histopathological features and attention-based aggregation, is built to facilitate predictions of progression-free survival after radical prostatectomy. Together, these models demonstrate the potential of several computer aided diagnosis tools, and pave the road for utilizing computational approaches to optimize and improve the efficiency of prostate cancer diagnosis and risk stratification.

The dissertation of Jiayun Li is approved.

Alex Anh-Tuan Bui

Benjamin M. Ellingson

Shyam Natarajan

Corey W. Arnold, Committee Co-Chair

Denise R. Aberle, Committee Co-Chair

University of California, Los Angeles

2021

Dedicated to my family and friends

Special gratitude to my beloved parents and husband

TABLE OF CONTENTS

1 Introduction **1**

1.1 Motivation 2

1.1.1 Semantic image segmentation for histopathological images 3

1.1.2 Classification and region of interest localization for whole slide images 4

1.1.3 Progression-free survival analysis with self-supervised learning 5

1.2 Contributions 6

1.3 Outline of the dissertation 8

2 Background **10**

2.1 Prostate cancer 10

2.1.1 Gleason grading system 11

2.1.2 Whole slide images 12

2.2 Computer aided diagnosis tools for whole slide images 13

2.2.1 Semantic image segmentation models for whole slide images 15

2.2.2 Semi-supervised semantic image segmentation for whole slide images 20

2.2.3 Classification and region of interest detection for whole slide images . 21

2.3 Computer aided progression model 27

2.3.1 Survival analysis 28

2.3.2 Progression-free survival analysis with pathomics features 30

3 Semantic Image Segmentation with Multi-scale Information: A Multi-scale U-Net Model **32**

3.1 Overview 32

3.2 Dataset 33

3.3 Method 33

3.3.1 Multi-scale U-Net 33

3.4 Experiment and results 39

3.4.1 Evaluation metrics 39

3.4.2 Results and discussion 40

3.5 Summary 44

3.6 Appendix 46

3.6.1 Additional experiments results 46

3.6.2 Color normalization for whole slide images 48

4 Semantic Image Segmentation with Weak Labels: An EM-based Semi-supervised segmentation model 52

4.1 Overview 52

4.2 Method 53

4.3 Experiment 57

4.3.1 Dataset and image preprocessing 57

4.3.2 Evaluation metric 59

4.3.3 Details on model implementation and training 59

4.4 Results 60

4.5 Discussion 61

4.6 Summary 65

5 Whole Slide Image Classification: A Multi-resolution Multiple Instance Learning Model 66

5.1 Overview 66

5.2 Method 67

5.2.1 Problem definition 67

5.2.2 Attention-based MIL with instance dropout 69

5.2.3 Attention-based tile selection 70

5.2.4 Multi-resolution WSI classification 71

5.3 Experiment 71

5.3.1 Dataset and data preprocessing 71

5.3.2 Implementation details 73

5.3.3 Evaluation metrics 74

5.3.4 Model visualization 75

5.3.5 Model comparison 76

5.4 Results 78

5.5 Discussion 85

5.6 Summary 87

5.7 Appendix 88

5.7.1 Detailed model architecture 88

5.7.2 Blue ratio conversion 88

5.7.3 K means clustering 88

5.7.4 Different CNN architectures for the feature extractor 92

5.7.5 Tissue region detection 92

6 Progression-free Analysis with Pathomic Features 95

6.1 Methods 96

6.1.1 Tumor region detection 96

6.1.2 Histopathological feature extraction 98

6.1.3 Aggregate tile-level features for progression free survival analysis . . . 101

6.2 Dataset 104

6.2.1 Data preprocessing 105

6.3 Experiment 106

6.3.1 Model training 106

6.3.2 Evaluation metrics 107

6.3.3 Comparison models 107

6.4 Results 109

6.5 Discussion 109

6.6 Summary 111

6.7 Appendix 112

6.7.1 Tumor detection stage model performance 112

7 Conclusion and Future Work 113

7.1 Summary of contributions 113

7.2 Future work 114

References 118

LIST OF FIGURES

1.1 Overview of deep learning methods developed in this dissertation to improve the diagnosis of prostate cancer. 8

2.1 Examples of various artifacts that exist in WSI datasets. (A) shows a slide with pen markers and dust. (B) presents a slide that contains many bubbles in both background and tissue regions. An example of tissue folding artifact is demonstrated in (C). Note these images are downsampled and cropped from the original WSIs for better visualization purposes. 14

2.2 Examples on different computer vision tasks. 1) Image classification models usually assign one or more labels to the whole image; 2) In addition to labels, image detection models also output bounding boxes that localize relevant objects; 3) The goal of semantic image segmentation is to output class labels for each pixel within the image. 16

2.3 An example of a CNN architecture. CNNs usually contain multiple different layers, such as convolutional layers, pooling layers, non-linear activation layers and fully connected layers. 18

2.4 An example of an FCN architecture. Different from CNNs, which output one or more global labels for the entire image, FCNs generate predictions for each pixel. Up-sampling convolution operations are utilized to recover spatial resolution of features maps. 19

2.5 Differences between fully supervised image classification models and MIL models in a binary classification scenario. Orange triangles denote positive samples and blue triangles represent negative samples. For fully supervised classification tasks, each input data is usually associated with one or more labels as shown in the top row. Yet, MIL models mainly deal with the scenario where a bag of images has one or more labels as demonstrated in the bottom row. Figures are better viewed in color. 24

3.1 Variations in gland size. (a) shows a tile with heterogeneous Gleason grades (G3, G4 and benign glands). Pathologist annotation mask is shown in (b). The high-grade cancer (G4) areas are shown in red, low-grade cancer (G3) areas are denoted as pink, benign glands are indicated by green, and stroma areas are represented by blue. These images demonstrate the heterogeneity of glands both between grades (*e.g.,* glands A and C) and within the same grade (*e.g.,* glands A and B). 35

3.2 Architecture of the multi-scale patch-based U-Net. The whole image was divided into multiple non-overlapping 100 × 100 sub-tiles. To capture contextual information, a 200 × 200 patch (framed in yellow) and a 400 × 400 patch (framed in black) were extracted around each centered 100 × 100 patch (framed in red). Features of different sizes were either down-sampled or up-sampled to 200 × 200, and concatenated into 64 × 200 × 200 feature maps that were input to a U-Net model. The final layer output a 4 × 100 × 100 probability map, each channel of which corresponded to a probability map of one class. 37

3.3 Segmentation masks generated by the U-Net and the multi-scale U-Net. Both ground truth masks and predictions are overlaid on original image tiles for easy interpretation. Colors follow the same schema illustrated in Figure 3.1. The first row shows segmentation results for an image tile with three tissue types (benign, stroma, and G3 cancer). The second row shows a representative image tile with three tissue types (stroma, G3 and G4 cancer). White arrows point to border areas that both models struggle with. 42

3.4 Segmentation results comparison for the multi-scale U-Net and the U-Net. Colors follow the same schema illustrated in Figure 3.1. The multi-scale U-Net successfully segmented the large irregular benign gland, while the U-Net with single scale input did not. 43

3.5 Segmentation results for two challenging tiles. Colors follow the same schema illustrated in Figure 3.1. The first row shows an image tile containing G4 cancer with poorly-formed glands. Glands were less differentiated on that tile, likely increasing segmentation difficulty. The second row presents a tile with a mixture of small high-grade glands and small low-grade glands. 43

3.6 Original histopathological images and normalized images with various methods: Reinhard [RAG01], Macenko [MNM09] and Vahadane [VPS16]. Images before normalization contain considerable variations as shown in the first row. Images are better viewed in color. 50

3.7 Additional examples on original and normalized histopathological images. 51

4.1 Differences between coarse contours and pixel-wise annotations. (A) A whole slide image with contour annotations visualized at 0.4x. Many tiles can be extracted from these contours. (B) A 1200 × 1200 tile sampled from one of the G3+G3 contours on A. It only has an image-level label of G3 inherited from the contour-level label. (C) A tile with pixel-wise annotations. The low-grade cancer (G3) and stromal areas are indicated by pink and blue colors respectively. (Figures are best viewed in color.) 54

4.2 Overview of EM-based semi-supervised semantic segmentation. LMAs are generated by pathologists so that enclosed regions only contain tissues of the designated label (*e.g.,* A 'G3+3' contour should contain purely G3 glands and stroma, devoid of benign glands or glands of other grades). EM-based algorithms are initialized with a multi-scale U-Net (as shown in B) trained on small amount of tiles with gland-level annotations, and trained on tiles with only image-level labels (as shown in A) extracted from LMAs on histopathological slides. In the E-step, the current model is applied to generate pixel-wise probability maps (as shown in C). To prevent the model from degeneracy, these probability maps are updated by a bias that has been adaptively selected by minimizing the KL divergence between the prior stroma versus epithelium distribution and the average model output distribution. Prediction masks (as shown in D) generated from the E-step are utilized to optimize model parameters in the M-step. To improve training, a small portion of patches with gland-level annotations are combined with patches with image-level labels in each batch. The EM-based method will iteratively update segmentation masks and model parameters until convergence. (Figures are best viewed in color.) 58

4.3 Segmentation masks for tiles in the test set. Stromal and benign areas are denoted by blue and green colors respectively. The high-grade (G4, G5) and low-grade (G3) cancer areas are represented by red and pink colors respectively. The first column shows that the initial model delivers inferior performance in segmenting epithelial areas, likely due to the small amount of available supervised training data. Both EM-based models (shown in the second and third columns) are able to improve segmentation performance using weakly labeled tiles. The best performance is achieved by adding 0.3 strongly labeled tiles during EM training. (Figures are best viewed in color.) 62

5.1 Overview of the proposed whole slide image detection and classification model. The model consists of two stages: a cancer detection stage at a low magnification and a cancer classification stage at a higher magnification for suspicious regions. Both stages contain a CNN feature extractor, which is trained in the MIL framework with slide-level labels. Specifically, the detection stage model is trained with all tiles extracted from slides at 5x to differentiate between benign and malignant slides. The attention module in the detection stage model produces a saliency map, which represents relative importance of each tile for predicting slide-level labels. Then we use the K-means clustering method to group tiles into clusters based on tile-level features. The number of tiles selected from each cluster is determined by the mean of cluster attention values. Discriminative tiles identified by the detection stage model are then extracted at 10x and fed into the classification stage model for cancer grade classification. 68

5.2 ROC and PR curves for detection stage models on our test set and external dataset. In the detection stage, models were trained to distinguish malignant and benign slides with all tiles extracted from slides at 5x. 79

5.3 WSIs overlaid with attention maps generated from the first stage cancer detection model. Pen markers as mentioned in §5.3.1 indicate cancerous regions. The first row shows attention maps from the model with instance dropout, while the second row is from the model without using instance dropout. Figures are best viewed in color. 80

5.4 Visualization of discriminative regions within tiles for TP and FP predictions. For each slide, we selected the top three tiles with the highest attention weights from the model, which were then forwarded to the model to generate activations and gradients for Grad-CAM. 81

5.5 t-SNE visualization of slide-level features. Black dots denote benign, purple dots indicate LG, and orange dots represent HG slides. (A) is the slide-level representations from the detection stage model. There is distinct separation between benign and cancerous slides. (B) shows the slide-level features from the classification stage model. We can see a better separation between LG and HG slides. 82

5.6 Confusion matrix for Gleason grade group prediction. 84

5.7 Br conversion. We performed Br conversion for slides at 5x. The first two rows demonstrate tiles from a benign slide and the bottom two show ones from a malignant slide. (A) are 3 tiles with the highest average tile-level Br values, and (B) are ones with the lowest Br values. We can see that the br conversion is able to highlight regions with most nuclei. 91

5.8 Examples of tissue detection results for biopsy slides. The processing pipeline is able to separate background and pen markers from tissue regions. This could help downstream models focus on the relevant tissue regions. 93

5.9 Additional examples of tissue detection results for prostatectomy slides. 94

6.1 Overview of the proposed PFS prediction system. The system mainly consists of two stages: 1) tumor detection stage, where a tile-level cancer classification model is applied to produce cancer probability map for each slide; 2) PFS prediction stage, where representations are extracted from selected tiles and aggregated into a slide-level feature vector for PFS prediction. The figure shows an example for self-supervised learning-based feature extraction with the MoCo model. Rather than using task specific labels, MoCo utilizes contrastive loss, which encourages smaller distances between features from augmented views of the same tile, but larger distance among other tiles. 97

6.2 Examples of cancer probability maps produced from the tile-level cancer classification model. Top tiles with highest cancer probabilities are selected for PFS prediction. 112

LIST OF TABLES

3.1 Model performances on segmenting prostate cancer (PCa), benign glands (BN) and stroma (ST). 41

3.2 Model performances on segmenting Gleason 4 (G4), Gleason 3 (G3), benign glands (BN), and stroma (ST). 41

3.3 Overall model segmentation performances. For each model, results from 5-fold cross validation were gathered into a overall confusion matrix. Evaluation metrics were then computed from the confusion matrix. 47

3.4 Average processing time for different stain normalization methods. 49

4.1 Model performances on segmenting stroma, high-grade (HG), low-grade (LG), and benign (BN) glands. 63

5.1 Number of slides for each Grade group. 72

5.2 Model performance on BN, LG, and HG slides classification. Cohen's Kappa and overall accuracy are reported in the table. To evaluate model performances on detecting malignant slides, probabilities for LG and HG are combined. AUROC and AP for cancer detection are also included in the table. 83

5.3 Details on models for experiments. 84

5.4 Cancer detection stage model architecture. 89

5.5 Cancer classification stage model architecture. 90

5.6 Model performance on BN, LG, and HG slides classification 92

6.1 Handcrafted features used for survival analysis. Tiles are first converted into HSV color space. H channel and the binary mask for tissue regions are fed into the Pyradiomics package is used to compute features. 101

6.2 Concordance index (c-index) of models using different feature pipelines and aggregation methods on predicting progression-free survival for prostate patients after radical prostatectomy. 108

ACKNOWLEDGMENTS

I would like to give my utmost appreciation and respect to my Ph.D advisor, Dr. Corey Arnold for his continuous support, inspiration and guidance throughout my Ph.D journey. His visions, knowledge, motivations have deeply inspired me to keep learning, improving and growing my skills in the scientific field. I want to offer my sincere gratitude to Dr. Denise Aberle. She inspired and encouraged me to pursue my studies and research in this interdisciplinary area. I also want to thank my other committee members: Dr. Alex Bui, Dr. Benjamin Ellingson, and Dr. Shyam Natarajan for their invaluable insights and feedbacks for this research.

I want to thank Dr. William Speier for providing constructive suggestions, discussions and help on this research. I'm extremely thankful to Dr. Anthony Sisk, Dr. William D. Wallace, and Dr. Huihui Ye for helping me understand essential clinical knowledge related to this dissertation, and providing labels for our research dataset. I also want to express my appreciation to Dr. Frank Meng, Dr. Ricky Taira, and Dr. William Hsu for their mentorship on various machine learning and informatics topics. I would like to thank Shawn Chen for helping set up servers and providing technical support to this research. I would like to thank Isabel Rippy and Denise Luna for helping me with administrative issues during my Ph.D study.

I would also like to acknowledge all current and past members of computational diagnostics (CDx) group and medical imaging informatics (MII) group for creating such a supportive and collaborative culture. Special thanks to my colleagues Wenyuan Li, Johnny Ho, Karthik Sarma, Zichen Wang, and Shiwen Shen, who have collaborated with me on many different projects and provided valuable suggestions on my paper writing and presentations. I want to thank Jennifer Polson, Harry Zhang, and Yiwen Meng for their help and support during my Ph.D journey.

In addition, I want to thank our collaborators from Cedars-Sinai medical center: Dr.

Arkadiusz Gertych, Dr. Beatrice Knudsen, Nathan Ing, and Zhaoxuan Ma for collaborations on the semantic image segmentation project and sharing us annotated datasets.

I would also like to thank UCLA for offering the Graduate Division Fellowship and the Bioengineering Fellowship for my study and research. This research is also supported by the UCLA Radiology Department Exploratory Research Grant Program (16-0003), NIH/NCI 5P50CA092131-15:R1 and NIH/NCI R21CA220352.

Last but not least, I would like to thank my friends and my family. They have provided emotional support and encouraged me to always stay strong and be optimistic especially during difficult times.

VITA

2011–2015	B.S. (Electronic and Information Science and Technology), Fudan University, Shanghai, China.
2015-2021	Graduate Student Researcher, Computational Diagnostic and Medical Imaging Informatics lab, UCLA, Los Angeles, California.
2018	Data scientist intern (summer), Ancestry.com
2019 & 2020	Software engineer intern in machine learning (summer), Google

PUBLICATIONS

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 2017 (Vol. 2017, p. 1140).

Li J, Speier W, Ho KC, Sarma KV, Gertych A, Knudsen BS, Arnold CW. An EM-based semi-supervised deep learning approach for semantic segmentation of histopathological images from radical prostatectomies. Computerized Medical Imaging and Graphics. 2018 Nov 1;69:125-33.

Li J, Li W, Gertych A, Knudsen BS, Speier W, Arnold CW. An attention-based multi-resolution model for prostate whole slide image classification and localization. In CVPR 2019 MVD Workshop.

Li J, Li W, Sisk A, Ye H, Wallace WD, Speier W, Arnold CW. A Multi-resolution Model for Histopathology Image Classification and Localization with Multiple Instance Learning. Computers in Biology and Medicine. 2021 Feb 10:104253.

Wang Z*, **Li J***, Pan Z, Li W, Sisk A, Ye H, Speier W, Arnold CW. Hierarchical graph pathomic network for progression free survival prediction. In submission.

Li W, **Li J**, Sarma KV, Ho KC, Shen S, Knudsen BS, Gertych A, Arnold CW. Path R-CNN for prostate cancer diagnosis and gleason grading of histological images. IEEE transactions on medical imaging. 2018 Oct 12;38(4):945-54.

Ing N, Ma Z, **Li J**, Salemi H, Arnold C, Knudsen BS, Gertych A. Semantic segmentation for prostate cancer grading by convolutional neural networks. In Medical Imaging 2018: Digital Pathology 2018 Mar 6 (Vol. 10581, p. 105811B).

Ebrahimpour MK, **Li J**, Yu YY, Reesee J, Moghtaderi A, Yang MH, Noelle DC. Ventral-dorsal neural networks: object detection via selective attention. In 2019 IEEE Winter Conference on Applications of Computer Vision (WACV) 2019 Jan 7 (pp. 986-994). IEEE.

Li W, Wang Z, **Li J**, Polson J, Speier W, Arnold CW. Semi-supervised learning based on generative adversarial network: a comparison between good GAN and bad GAN approach. In CVPR Workshops 2019 May 16.

Li W, Wang Z, Yue Y, **Li J**, Speier W, Zhou M, Arnold C. Semi-supervised learning using adversarial training with good and bad samples. Machine Vision and Applications. 2020 Sep;31(6):1-1.

CHAPTER 1

Introduction

Prostate cancer accounts for nearly 20% of new cancer diagnosed in men, and is the most prevalent and second deadliest non-skin cancer in men in the United States [SMJ19]. About one man in nine will be diagnosed with prostate cancer during his lifetime and about one in 41 will die of their disease [QBD13]. Most men are diagnosed with indolent disease that do not require immediate definitive local therapy (*i.e.,* surgery or radiotherapy). To risk-stratify these patients for management, three clinical measurements are routinely used: pre-treatment serum abundances of prostate specific antigen (PSA), tumor size and extent as estimated by DRE or sometimes multi-parametric MRI (mpMRI), and tumor grade, as quantified by International Society of Urological Pathology (ISUP) Grade Groups (and formerly via the closely related Gleason score) [EZS16]. These three simple clinical features are used to stratify newly-diagnosed prostate cancers into low-, intermediate- and high-risk cases. Patients are treated differently depending on their risk groups, with low-risk patients typically receiving active surveillance (AS) while high-risk patients receive definitive local therapy (surgery or radiotherapy at equipoise), often with adjuvant hormone therapy. Interpretation and grading of histopathological slides plays an essential role in treatment planning for prostate cancer patients. Histopathological evaluation of prostate cancer mostly relies on pathologists to qualitatively summarize heterogeneous histological growth patterns with the Gleason grading system.

Currently, Gleason grading system is the current best method for prostate cancer diagnosis and is a critical component in clinical survival assessment and treatment planning

[EZS16]. However, the process of looking through the large tissue slide for suspicious areas, which potentially only occupy a small percentage of the entire surface area, can be tedious and time-consuming. Studies have also shown that the diagnosis of Gleason grades can be prone to inter- and intra-observer disagreements [LD12, HKZ14, OSY96, SM85]. Moreover, risk stratification based on Gleason grading system may be imperfect. For example, up to a third of men with intermediate-risk disease will suffer a relapse, indicative of under-treatment. The Prostate Testing for Cancer and Treatment (ProtecT) trial found no significant difference in mortality at 10 years between patients on AS and those who underwent immediate surgery [HDL16].

1.1 Motivation

The aforementioned challenges underscore the need of computer aided diagnosis (CAD) tools that could assist pathologists to localize suspicious regions, serve as a second reader to potentially improve rater agreements, as well as extract and effectively aggregate underlying representations from digitized histopathological slides for better risk stratification.

The recent development of digital whole slide scanners has enabled traditional glass tissue slides to be scanned into high resolution images (also referred as whole slide images), which has led to a new wave of image analysis research on this new source of "big data". Machine and deep learning methods have demonstrated promising results in many medical computer vision tasks such as classification, tumor detection and segmentation [BSR19, CGZ20, SCM20].

Yet, there are unique challenges in analyzing these whole slide images (WSIs): 1) the size of WSIs is enormous (*i.e.,* each WSI may contain over billions pixels); 2) tissue contents within a WSI are highly heterogeneous; 3) available labels for WSIs are usually at global levels and sparse (*e.g.,* one label for billions of pixels) and how to aggregate local information for global predictions is challenging; 4) artifacts such as stain variations, tissue folding, and pen markers appear frequently in WSIs.

1.1.1 Semantic image segmentation for histopathological images

Previous studies have been done in developing automated Gleason grading systems to help improve diagnostic accuracy. A commonly used approach is to extract handcrafted features and apply classifiers on pre-selected small image tiles, each of which only contains one tissue class [FSJ07, DFT10, NSJ12]. Yet, the requirement of pre-extracted image tiles with homogeneous tissue content precludes their utility as computerized tools for much larger and more heterogeneous WSIs. Semantic image segmentation methods, in contrast, could provide dense predictions for each pixel and estimate percentage distribution of different tissue classes, which could also be used as a pre-step for quantitative histopathological features extraction.

Training fully convolutional neural networks for histopathological images could be challenging due to the high resolution nature of these images, limited available memory of a graphics processing unit (GPU) and limited amount of annotated samples. Two approaches to handling this challenge are resolution downsampling and patch extraction. In downsampling, high resolution images are scaled down to more manageable sizes, at the cost of the loss of potentially discriminative fine details. In patch extraction, images are divided into (possibly overlapping) sub-patches that are then treated as independent training samples. This approach allows for the analysis of full resolution data, but may lead to the loss of context information. Moreover, both morphological patterns of glands and fine-grained features of nuclei are useful in prostate histopathological image diagnosis. In this dissertation, a multi-scale U-Net model is developed to leverage representation from different scales to facilitate pixel-wise Gleason grade prediction for histopathological images with heterogeneous tissues. This may enable more precise estimation of percent of different Gleason patterns, which are shown to be related to long-term outcomes [CDL07, CMS16].

Semantic segmentation models demonstrated promising results in generating dense local predictions of Gleason grades. However, training image segmentation models requires large amount of samples annotated at pixel-level, which are expensive to obtain especially in the medical domain. This may limit the performance and the transferability of segmentation models. Conversely, image-level labels derived from region annotations can be easily generated. Thus, leveraging information from large-scale weakly labeled dataset can be an alternative way to improve semantic image segmentation. Learning with image-level labels for segmentation can be challenging, since the relationship between labels and predictions becomes many-to-one rather than one-to-one. Some previous work utilized pre-defined functions such as maximum pooling and mean pooling to combine pixel-wise predictions, which were generated from the fully convolutional networks, to an image-level prediction. Loss was then backpropagated from image-level labels. In this dissertation, pixel-wise labels are considered as hidden variables and an expectation maximization (EM)-based semi-supervised image segmentation approach, which exploits weakly annotated samples, is developed to further improve the performance of the multi-scale U-Net model.

1.1.2 Classification and region of interest localization for whole slide images

Semantic segmentation models and tile-level classification models could provide predictions at local levels. Yet, the diagnosis of WSIs is usually done at a global level. Therefore, how to effectively summarize local predictions into a slide-level output that can potentially be used as a second reader during histological evaluation remains a challenge.

In addition, though the size WSIs is very large, suspicious regions may only take up a small portion of it with other regions being background, stroma, and benign tissues. Currently, pathologists need to scan through a histopathological slide at a relative low magnification (*e.g.,* 4x), searching for relevant regions on which to zoom in at a higher magnification (*e.g.,* 10x) and ascertain Gleason scores. Pathologists may need to zoom in and out several times to evaluate and grade multiple suspicious regions. Therefore, a CAD tool, which could localize areas of interest and provide slide-level predictions, can potentially save significant amount of time on screening benign regions or cases and looking for cancerous areas. Unlike most previous work that relied on labor intensive labels or utilized fixed aggregation functions, this dissertation presents a multi-resolution multiple instance learning model for WSI classification and localization. The model follows the diagnosis workflow by pathologists, which first detects cancerous areas and then zooms in on the suspicious regions to make grade prediction.

1.1.3 Progression-free survival analysis with self-supervised learning

Gleason grading system, which is commonly used by pathologists to evaluate histopathological slides of prostate cancer patients, is a strong predictor for prostate cancer prognosis. Yet, the grading system describes diverse tumor histology with fixed numerical categories, and may not be able to model the complex biological signals and capture the full spectrum of underlying various histopathological patterns. Also determining Gleason grades remains a process that is relatively subjective and has been shown to have low inter-observer agreement across pathologists [LD12, HKZ14, OSY96, SM85]. Development of CAD tools that can provide second opinions during diagnosis is one possible way to alleviate these challenges by improving the efficiency and reproducibility of the pathology evaluation. Computer aided progression systems, on the other hand, could leverage quantitative features to measure underlying tumor characteristics and further facilitate better risk stratification. Deep learning method, when trained with sufficient amount of data, can be a powerful tool for extracting discriminative visual and sub-visual representations from images. Nevertheless, obtaining a large-scale dataset with outcome labels is non-trivial. Using hand-crafted features generated from pre-selected regions of interest has been widely utilized in previous work for prostate cancer progression prediction. This dissertation exploits a self-supervised learning-based method to extract informative features from suspicious tiles and investigates different aggregation methods, such as attention-based and graph convolutional neural network-based approaches, to combine tile-level features for case-level progression-free survival prediction.

1.2 Contributions

In a nutshell, in order to address aforementioned challenges, two types of deep learning-based CAD tools have been developed in this work to assist the pathological evaluation of WSIs for prostate cancer. To further facilitate characterization of heterogeneous histopathological patterns embedded in the WSIs, a deep learning system that incorporates self-supervised learning-based representations and trainable aggregation methods is developed for progression-free survival prediction.

Main contributions of this dissertation can be summarized in the following specific aims:

Aim1. Develop fully supervised multi-scale semantic segmentation models that produce pixel-wise Gleason grading, using tiles from whole mount prostatectomies and corresponding manual annotations from pathologists.

a. To investigate different color normalization algorithms to account for color variations in WSIs due to inconsistencies in raw material preparation, staining protocols, scanning condition, and *etc.*

b. To develop computational algorithms to produce dense predictions on large high-resolution histopathological images.

c. To improve segmentation performance with multi-scale architectures such as U-Net, which can extract deep representations both from nuclei and glandular structures.

Aim2. Develop semi-supervised approaches to leverage the full potential of large-scale weakly-labeled clinical data for refining and further improving visual semantics learned under Aim 1.

a. To develop an EM-based semi-supervised semantic segmentation model with a small set of fully annotated data and a large set of weakly annotated data.

b. To improve the semi-supervised learning by incorporating prior knowledge about epithelium-stroma distribution as bias into EM training.

Aim3. Develop a CAD tool for whole slide image analysis that could localize suspicious regions and produce the slide-level Gleason grade prediction, using slides from a large-scale non-curated prostate biopsy dataset.

a. To develop a multiple instance learning-based model that can be trained with slide-level labels instead of relying on fine-grained annotations at pixel or region level.

b. To develop a multiple resolution model that first processes the entire slide at relatively low magnification to detect potentially cancerous regions and then zooms in on these areas for Gleason grading.

c. To investigate the effectiveness of attention-based informative tiles selection method and visualize the learned model and features.

Aim4. Develop a progression-free survival prediction model to better characterize diverse histopathological representations embedded in the WSIs by incorporating self-supervised learning features and aggregation strategies.

a. To develop a system that can identify suspicious regions and leverage rich prognostic information embedded in WSIs for better progression-free survival prediction.

b. To investigate the effectiveness of deep representations learned from a self-supervised learning model for capturing underlying histopathological patterns.

c. To investigate different aggregation methods including attention-based and graph convolutional neural network-based approaches for combining local information into global representations.

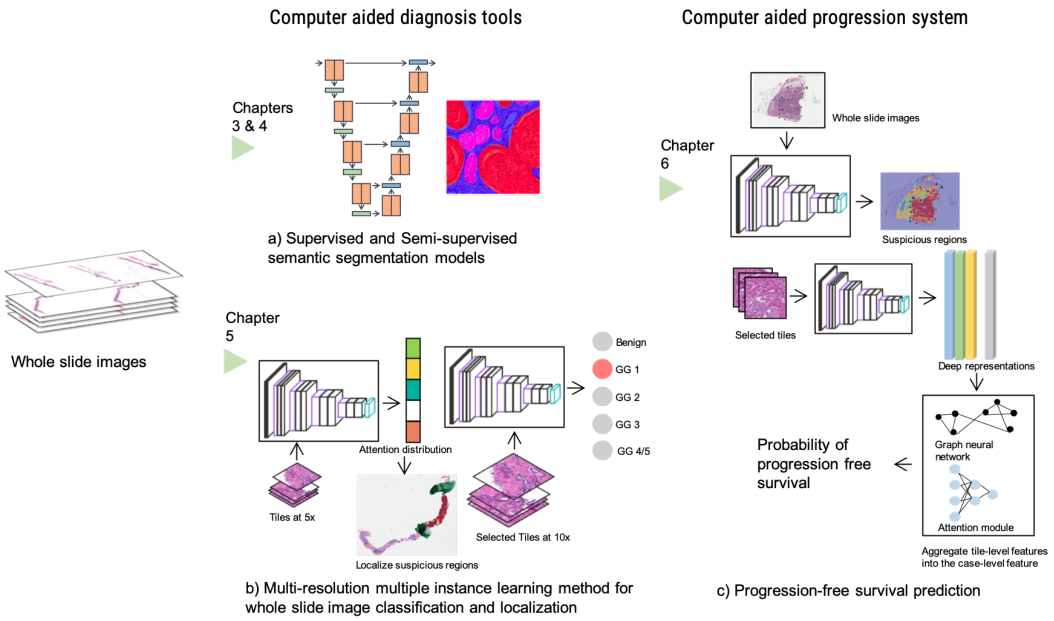
Computer aided diagnosis tools
Computer aided progression system
Chapters 3 & 4
a) Supervised and Semi-supervised semantic segmentation models
Whole slide images
Chapter 5
Attention distribution
Tiles at 5x
Localize suspicious regions
Selected Tiles at 10x
Benign
GG 1
GG 2
GG 3
GG 4/5
b) Multi-resolution multiple instance learning method for whole slide image classification and localization
Chapter 6
Whole slide images
Suspicious regions
Selected tiles
Deep representations
Graph neural network
Attention module
Probability of progression free survival
Aggregate tile-level features into the case-level feature
c) Progression-free survival prediction

Figure 1.1: Overview of deep learning methods developed in this dissertation to improve the diagnosis of prostate cancer.

1.3 Outline of the dissertation

Figure 1.1 presents a diagram of core models developed in this research, including two CAD tools for semantic image segmentation and slide classification, and a deep learning model for progression survival prediction. This dissertation is organized as follows:

Chapter 2 provides the background on prostate cancer diagnosis, risk stratification, whole slide images, supervised and semi-supervised deep learning methods for semantic image segmentation, multiple instance learning framework, survival analysis, and selected review of related work.

Chapter 3 presents a novel multi-scale U-Net model for semantic image segmentation of histopathological images from prostatectomy slides. The model can deal with images with heterogeneous tissue contents and produce pixel-wise Gleason grading.

Chapter 4 extends the fully-supervised segmentation models with an EM-based semi-supervised learning framework, which enables training of a segmentation model with samples labeled at image-level. This leverages information contained in weakly-labeled dataset and alleviates the need of expensive fine-grained annotations for segmentation models.

Chapter 5 demonstrates a multi-resolution multiple instance learning-based model for WSIs classification and localization. The model addresses several challenges in analyzing the entire WSI and can be trained with slide-level labels.

Chapter 6 builds a deep learning system for progression-free survival prediction of prostate cancer patients. The proposed pipeline detects suspicious areas, extracts discriminative representations from selected regions, and aggregates tile-level information for case-level survival predictions.

Chapter 7 summarizes results of this dissertation, discusses limitations and potential future directions.

CHAPTER 2

Background

2.1 Prostate cancer

Prostate cancer is the most common and second deadliest non-skin cancer in American men [SMJ19]. Most prostate cancer cases are diagnosed while the tumor is still localized to the gland. Patients are stratified into low-, intermediate- and high-risk groups based on clinical measurements: pre-treatment serum abundances of prostate specific antigen (PSA), tumor size and extent as estimated by DRE or sometimes multi-parametric MRI (mpMRI), and tumor grade obtained from transrectal ultrasound-guided (TRUS) biopsies [MBB10]. Low-risk patients typically monitored by active surveillance (AS), while patients in the high-risk group may be intervened with radiotherapy and radical prostatectomy, with or without hormonal therapy.

However, patients within the same risk group may still have heterogeneous prognosis, and the over-diagnosis and over-treatment of clinically insignificant prostate cancer are also challenges for patient management [SA12, PPP15, LBN14]. Over-diagnosis is the detection of cancer, mostly through screening, that is asymptomatic, non-growing or slow growing that would not benefit from treatment and would not result in cancer-related death even if untreated. Over-treatment largely occurs because health care providers cannot definitively discriminate non-aggressive (low-risk) and aggressive (high-risk) cancer[SA12]. As a result of over-treatment, men receive unnecessary interventions with potentially severe side effects, including erectile dysfunction, urinary incontinence, and infertility [KZL10]. Although over-diagnosis will occur with any disease that is detected by screening, the over-diagnosis rate for prostate cancer is high and increases with age [SA12]. The 2018 Evidence Report and Systematic Review for the US Preventive Services Task Force found that over-diagnosis was as high as 50.4% in tumors detected as a result of screening, with 27 men needing to be diagnosed and potentially treated to prevent one prostate cancer death [FWD18]. Additionally, over-diagnosis and over-treatment of low-risk cancers lead to significant healthcare spending, with an estimated cost of $5,227,306 to prevent one prostate cancer death [SA11].

2.1.1 Gleason grading system

Histopathologic assessment is a key component for diagnosis of many diseases including prostate cancer. The evaluation mainly relies on pathologists using visual classification system to qualitatively describe diverse tumor histology. Gleason grading system is one such system that is commonly used to measure tumor growth patterns for prostate cancer. The Gleason grading system consists of five different histologic patterns from Gleason 1 (G1) indicating tissue that closely resembles normal prostate glands to Gleason 5 (G5) representing undifferentiated carcinoma and exhibiting the highest risk for dissemination. The final Gleason score (GS) is generated by summing the most (primary Gleason pattern) and second most (secondary Gleason pattern) prevalent patterns in the tissue section.

A recent study proposed to revise the Gleason grading system with 5 Gleason Grade groups (GGs) to reduce the over-treatment of low-grade prostate cancer [EZS16]: GG 1 (GS ≤ 6), GG 2 (G3 + G4), GG 3 (G4 + G3), GG 4 (GS = 8) and GG 5 (GS ≥ 9). Patients with intermediate- to high-risk localized prostate cancer (GG ≥ 2) may be intervened with radiotherapy and radical prostatectomy, with or without hormonal therapy.

Currently, the diagnosis of prostate cancer relies on pathologists to examine multiple levels of biopsy cores at the scanning magnification, and identify suspicious regions for high power examination and immunohistochemistry if necessary. This process can be tedious and time-consuming. More importantly, some patterns, *e.g.,* ill-defined G4 versus tangentially sectioned G3, are prone to inter- and intra-observer variability. Therefore, the current clinical practice can be improved by computer aided diagnosis (CAD) tools that can function as primary screening, to localize suspicious regions, and be utilized as a second reader for Gleason grading.

2.1.2 Whole slide images

With the recent development of digital whole slide scanners with faster scanning speed and higher image quality, an increasing number of traditional glass tissue slides are being digitized into high-resolution slides. These digitized slides are referred as whole slide images (WSIs) or histopathological slides. The Food and Drug Administrations (FDA) have approved the use of WSIs for primary diagnosis [Gar16, HRS19, HPE15, NCJ20]. This facilitates the wider adoption of digital slides in routine clinical workflows and opens new research opportunities for analyzing this new type of data with various machine learning models including segmentation, detection, classification and *etc* [BSR19, SCM20].

Different from natural images, which usually have small to moderate sizes (*e.g.,* the size of CIFAR-10 images [KH09] is 32×32 and the average resolution for images in the ImageNet dataset [DDS09] is 469×387), WSIs could contain over a billion of pixels. This makes it almost impossible to forward the entire slide at high scanning magnification into GPUs. Moreover, though slide labels can be easily extracted from pathology reports, contents within a slide can be highly heterogeneous (*e.g.,* a slide labeled as G3+G4 could contain G3 glands, G4 glands, benign glands, stroma and *etc.*). In addition, preparation procedures, scanning protocols and conditions could cause many variations and artifacts such as stain variations, pen markers, dust, air bubbles, tissue folding, tissue tear, tangential cut and *etc* [TSS18, RPA13], which may potentially hamper the performance of downstream image analysis algorithms. The enormous image size, the heterogeneity of tissues and various possible artifacts create unique challenges on developing machine learning models for WSIs.

Some examples of artifacts of WSIs are shown in Figure 2.1. The preprocessing pipeline including tissue detection and stain normalization are implemented in this work to reduce some of aforementioned artifacts. Details on stain normalization are discussed in §3.6.2 and §5.7.5 presents details on the tissue detection pipeline.

2.2 Computer aided diagnosis tools for whole slide images

CAD tools for digital pathology are systems that are designed to assist pathologists to interpret slides. These tools usually take digitized slides or selected regions as inputs and output quantitative measurements such as histologic grading, cell counting and *etc* based on underlying visual patterns. CAD tools could potentially 1) help reduce the diagnosis time (*e.g.,* a CAD tool, which can detect regions of interests, will help pathologists quickly find diagnostic relevant regions and reduce the time spent on benign areas.); 2) operate as a second reader and improve the reproducibility and consistency of pathology diagnosis; 3) retrieve and search for cases with similar histologic patterns (*e.g.,* content-based image retrieval systems).

With the increasing availability of digitized slides, development of CAD tools is becoming an active research area and many tools have been developed for various disease types. For example, Bejnordi *et al* evaluated algorithms for automated detection and classification of lymph node metastases for breast cancer patients as part of the CAMELYON16 challenge [BVV17]. Top algorithms achieved comparable performances as pathologists' diagnosis without time constraint. Steiner *et al* evaluated the potential impact of CAD tools on helping pathologists detecting breast cancer metastasis [SML18]. The study showed that the CAD tool significantly increased the sensitivity of detecting micrometastases and greatly reduced the average review time per slide [SML18]. Hekler *et al* utilized a ResNet50 model to classify cropped image sections from WSIs of melanomas and nevi [HUE19a]. The model significantly outperformed the 11 pathologists regarding overall accuracy, sensitivity and specificity [HUE19a, HUE19b]. A two-stage deep learning system developed by Nagpal *et al*

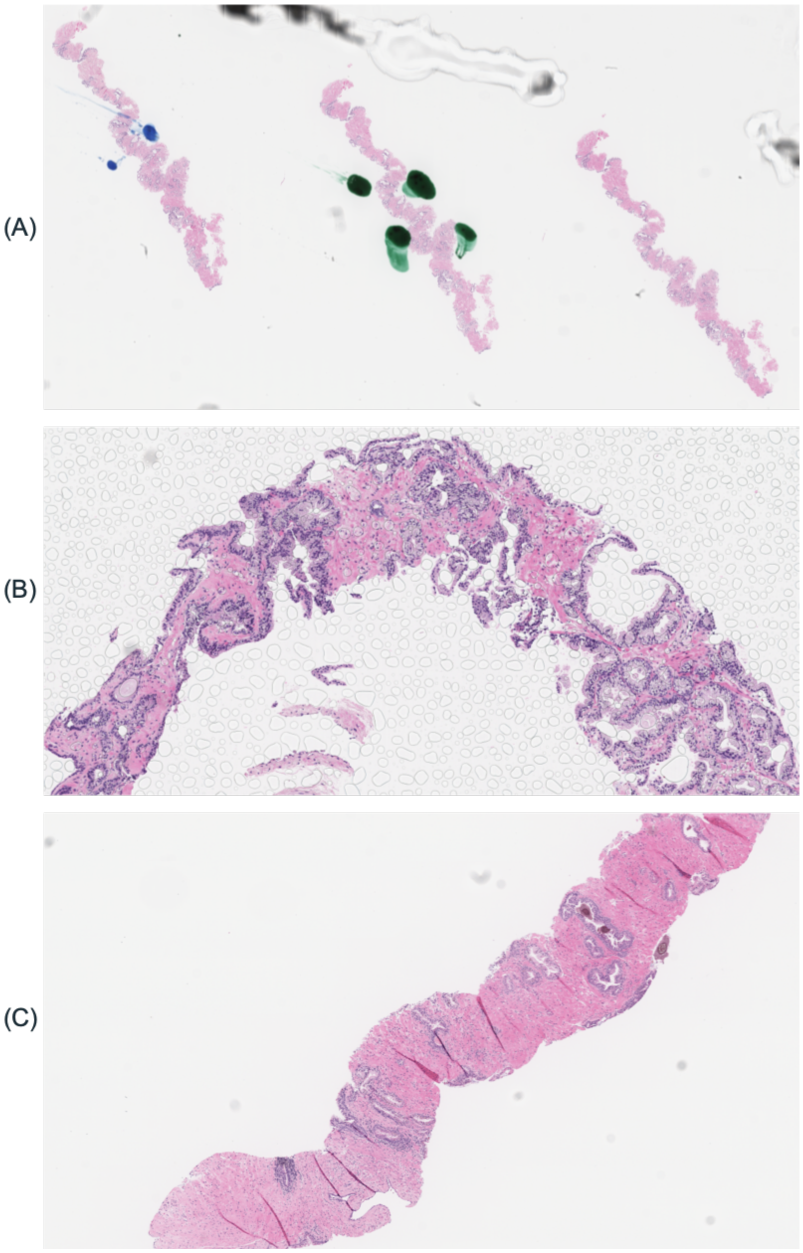

Figure 2.1: Examples of various artifacts that exist in WSI datasets. (A) shows a slide with pen markers and dust. (B) presents a slide that contains many bubbles in both background and tissue regions. An example of tissue folding artifact is demonstrated in (C). Note these images are downsampled and cropped from the original WSIs for better visualization purposes.

was utilized to provide automated Gleason grading for prostatectomy slides [NFL19, NFT20]. Besides detecting regions and providing automated diagnosis, CADs may improve clinical decision making by helping retrieve similar cases for references during diagnosis. Hegde *et al* proposed a similar image search (SMILY) system for histopathological data, which utilized embeddings generated from a deep ranking network to create a condensed patch representation for retrieval [HHL19, WSL14]. The system included refinement tools that enabled query with selected regions, query with example and query with concept [CRH19, HHL19]. Future evaluation conducted by [CRH19] demonstrated that the system could potentially open the "black-box" for deep learning models and increase the user trust in algorithms.

Two types of CAD tools for prostate WSIs are mainly developed in this dissertation: a tile-level segmentation tool, which could provide pixel-wise Gleason grade prediction and facilitate the estimation of percentage of different cancer grades; a slide-level classification and localization tool, which can highlight suspicious regions on large-scale images and be utilized as a second reader for Gleason grading.

2.2.1 Semantic image segmentation models for whole slide images

Different from image classification, which classifies the entire image into one or more classes, or object detection that localizes objects of interest, semantic image segmentation is a computer vision task that focuses on classifying each pixel into a certain class. Figure 2.2 demonstrates differences between various computer vision tasks.

2.2.1.1 Convolutional neural network

Convolutional neural networks (CNN) is among the most promising and widely used architectures for many computer vision tasks. After the CNN was initially proposed by Fukushima [FM82], weights sharing and gradient back propagation based training strategies were introduced by [WHH89]. LeCun *et al* developed the LeNet CNN architecture for document

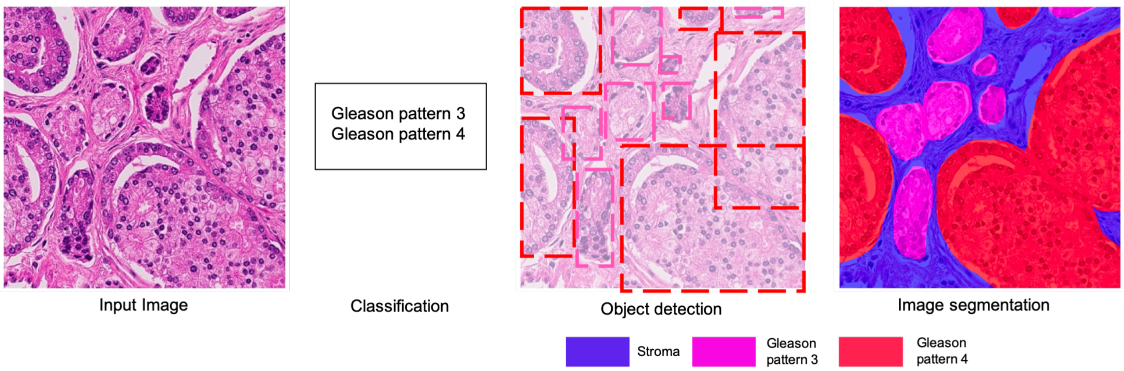

Figure 2.2: Examples on different computer vision tasks. 1) Image classification models usually assign one or more labels to the whole image; 2) In addition to labels, image detection models also output bounding boxes that localize relevant objects; 3) The goal of semantic image segmentation is to output class labels for each pixel within the image.

recognition [LBB98].

CNNs, which can extract hierarchical representations from images, mainly consist of three different types of layers: convolutional layers, pooling layers and non-linear activation layers. Feature maps are flattened and forwarded into fully connected layers for class prediction. Figure 2.3 shows an example for a CNN.

Convolutional layers contain parameterized kernels, also known as filters, which convolve with inputs to extract features. Local connectivity and weight sharing are two important characteristics of filters. Specifically, local connectivity refers to the concept that each unit (*i.e.,* neuron) is only connected to a small region (*i.e.,* receptive field) of the input feature map or image [LBB98, LBH15]. Weights of filters within a feature map are shared across different spatial locations, which represents the concept of weight sharing [WHH89, LBB98, LBH15]. These design paradigms effectively reduce the number of model parameters and retain the power of feature learning by exploiting the observation that pixels can be highly correlated with neighbors and local patterns are invariant to spatial locations.

Non-linear layers apply a non-linear activation function on feature maps, which enables the network to learn complex non-linear functions. Most commonly used non-linear activation functions include Sigmoid, Tanh, ReLU [Aga18] and LeakyReLU [XWC15].

Pooling layers reduce the spatial size of feature maps by aggregating representations within a small region into one. A unit in a pooling layer connects to a local area of the previous feature map and summarizes features with some statistical measures such as mean and max. Therefore, the pooling layer not only reduces the number of parameters required by the following layers but also improves the robustness of the network to small shifting.

Fully connected layers are usually incorporated at the end of the CNN and used to produce class scores. Different from convolutional layers where each neuron only connects to a small region, neurons in fully connected layers connect with all neurons in the previous layer.

In addition, batch normalization [IS15], instance normalization [UVL16], dropout [SHK14] and *etc* are also widely used layers that are developed to prevent model overfitting and facilitate model training. By stacking various layers, CNNs are powerful models for learning hierarchical representations from images. The deeper layers can connect to increasingly larger receptive fields and are able to learn high-level and class specific features, while shallower layers mostly extract low-level patterns such as edge and texture features. Some most commonly used CNN architectures include AlexNet [KSH17], VGG [SZ14], ResNet [HZR16], DenseNet [HLV17], EfficientNet [TL19], and MobileNet [HZC17].

2.2.1.2 Fully convolutional neural network

CNNs demonstrate significantly improved performances on image classification, detection and *etc*. Different from the image classification, image segmentation requires dense predictions for each pixel. Applying the CNN in a sliding window manner to produce predictions for centered pixels is a direct way to adopt CNNs for the segmentation task [CGG12, HAG14, GGA14]. Yet, this method is not efficient since it requires multiple passes through CNNs.

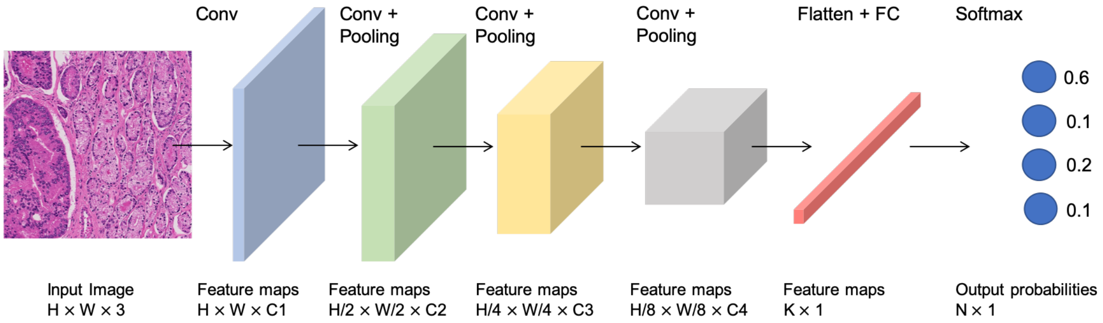

Figure 2.3: An example of a CNN architecture. CNNs usually contain multiple different layers, such as convolutional layers, pooling layers, non-linear activation layers and fully connected layers.

In CNNs, feature maps from the last convolutional layer are flattened and fed into the fully connected layers with fixed number of neurons. This discards spatial information of features, which is important in producing dense pixel-wise predictions. To keep spatial relationships, a fully connected layer can be converted into a convolutional layer by using convolutions with filter size equal to the feature map. For instance, instead of having a output vector of size $C \times 1$ to represent probabilities for C classes, fully convolutional neural networks (FCNs) output a score map of size $C \times k \times k$, which represents class probabilities at corresponding spatial locations.

Stacking of convolutional and pooling layers effectively increases receptive fields of filters without inducing too much computational cost. However, these operations reduce the spatial resolution of the input image (*i.e.*, k may be much smaller than the input image). Direct up-sampling the reduced size score map may lead to coarse segmentation maps [LSD15, SLD17]. Long *et al* proposed [LSD15, SLD17] to use upsampling convolution, also known as deconvolution or transposed convolution, to recover the score map resolution. In practice, this can be achieved by using convolution with fractional stride. The transposed convolution utilizes parameterized filters to learn how to up-sample the input feature map, and it can be considered as a reverse operation of the convolution. Figure 2.4 demonstrates an example of

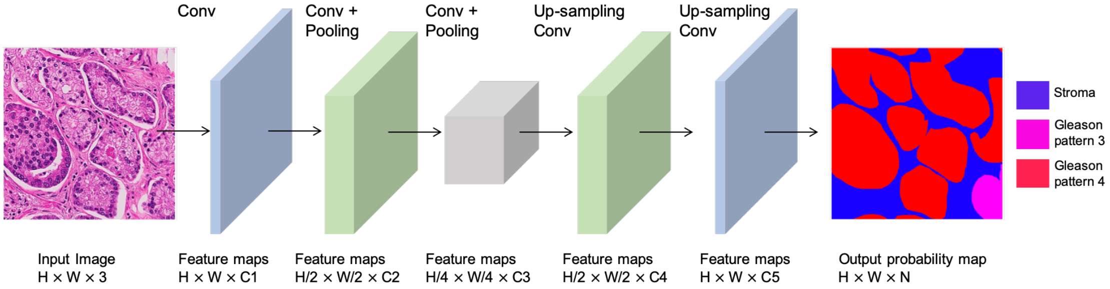

Figure 2.4: An example of an FCN architecture. Different from CNNs, which output one or more global labels for the entire image, FCNs generate predictions for each pixel. Up-sampling convolution operations are utilized to recover spatial resolution of features maps.

FCN with up-sampling convolutions.

To further refine the segmentation output, Long *et al* exploited skip connections that fused features from shallower layers with those from deeper layers [LSD15, SLD17]. Shallow layer features mostly capture fine-grained local patterns, while deeper layers generally extract global features. Combining features from shallower and deeper layers enables the models to produce local predictions in the context of global structures. The encoder-decoder based architecture is another type of deep segmentation models, which consists of symmetric contracting and expansion paths [NHH15, BKC17, RFB15]. For example, SegNet proposed by Badrinarayanan [BKC17] utilized an encoder with topologically identical structure as VGG16 and a decoder that used pooling indices for up-sampling [BKC17]. U-Net is another encoder-decoder based architecture that was designed for segmenting microscopy images [RFB15]. Data augmentations were excessively used to deal with limited training data, and it significantly outperformed other models in the ISBI cell tracking challenge 2015 [RFB15]. U-Net has been utilized and extended in various biomedical image segmentation tasks [MNA16, MBP20, DZY20]. R2U-Net, which replaced the feedforward convolutional layers with recurrent residual modules, was developed for end-to-end nuclei segmentation [AYT18]. Mehta *et al* extended the U-Net and developed a Y-Net for segmentation of breast biopsy images, which included a separate branch for classification [MMB18]. Y-Net was conceptually similar to the Mask R-CNN model [HGD17], but didn't require a region proposal network or instance-level annotations.

In this work, we further extend the U-Net model by explicitly exploiting multi-scale information for segmentation of prostate histopathological images. Details on the multi-scale U-Net model are included in Chapter 3.

2.2.2 Semi-supervised semantic image segmentation for whole slide images

Previous methods have achieved significant performances on semantic segmentation with FCNs. However, these models require pixel-wise labels. A large-scale dataset with expert annotations at gland-level or pixel-level is both time-consuming and expensive to obtain, especially in the medical domain. Several existing approaches have been developed to address the challenge of leveraging information embedded in data with weak annotations such as image tags, and bounding boxes [DTC16, KBF16, PC15, XZE12, XZE14, WYH15]. Learning with weak labels is often formulated in the multiple instance learning (MIL) framework [ATH02, DLL97, JHE17, XZE14] where training data consists of labeled bags with multiple unlabeled instances, with the goal to predict labels of unseen bags or instances. Noisy-OR [ZPV06], Generalized Mean (GM) [ZPV06], and log-sum-exponential (LSE) [RD00] are some commonly utilized methods to aggregate pixel-level probabilities into image-level prediction. For example, Pathak *et al* used the maximum pooling to combine predictions from heatmaps generated by the FCN for each class [PSL14]. Jia *et al* developed a constrained weakly supervised FCN model, which utilized the GM function to aggregate pixel-wise probabilities into image-level labels in order to segment cancerous areas on histopathological images of colon cancer [JHE17]. An aggregation function can be easily incorporated into an FCN network, but training errors are propagated through pixels with large prediction confidence, which can be affected by few significantly misclassified pixels [HSK16]. Thus, this type of method could be sensitive to initialization [PSL14].

EM-based weakly-supervised approaches usually consider pixel-wise labels as hidden variables and find the optimal solution by iteratively updating prediction masks and optimizing model parameters. Papandreou *et al* trained an EM model and employed a bias on model output to encourage at least ρ percentage of each image to be assigned to foreground [PCM15]. Those approaches require initialization with pre-trained models on the large ImageNet dataset, and tuned with a weakly labeled dataset for semantic segmentation. Different from their approaches, the EM-based segmentation model developed in this dissertation starts with an undertrained model, and leverages new information embedded in a weakly labeled dataset to improve the segmentation performance. The proposed EM-based approach is regularized by an estimated prior distribution, and significantly improves the segmentation performance of the initial fully supervised model.

2.2.3 Classification and region of interest detection for whole slide images

Another type of CAD tools developed in this dissertation focuses on classifying the entire whole slide image and identifying regions of interest (ROIs). This type of model could potentially function as primary screening, to localize suspicious regions, and be utilized as a second reader for Gleason grading.

Classification of small homogeneous ROIs pre-selected by pathologists has been the main focus of most early work in WSI image classification [FSJ07, DFT10, NSJ12]. Farjam *et al* developed a method to segment prostate glands with texture-based features, and then used the size and shape features of glands to classify image tiles into benign or malignant glands [FSJ07]. Nguyen *et al* used structural features of prostate glands to classify pre-extracted ROIs into benign, G3, and G4, achieving an overall accuracy of 85.6% [NSJ12]. In the work by Gorelick *et al*, a two stage Adaboost model was applied to classify around 991 sub-images extracted from 50 whole-mount sections of 15 patients [GVG13]. They achieved 85% accuracy for distinguishing high-grade (G4) cancer from low-grade cancer (G3).

However, the above algorithms require a set of pre-extracted image tiles with homogeneous tissue contents, which may not be generalizable to larger and more heterogeneous images. Moreover, accurate localization of such small image tiles is a non-trivial problem [DFT10]. Rather than attempting to classify the entire image tile, some work has addressed this challenge by developing segmentation models that can provide pixel-wise predictions for tiles with various tissue contents [GIM15, LSH18, LSH17, IML18, LLS18]. Li *et al* developed a novel region-based segmentation model (*i.e.*, Path RCNN), with an epithelial network head and a grading network for multi-task prediction [LLS18]. The model achieved the state-of-the-art performance in Gleason pattern segmentation.

However, these models still analyzed tiles instead of the entire slide. With an increasing number of scanned slides and computing power, research in WSI has been shifting to slide-level analysis [NFL19, NFT20, NKG19]. Litjens *et al* developed a deep convolutional neural network (CNN)-based model, which classified patches into cancer and non-cancer, and then predicted slide-level labels by applying the patch-level classifier to every pixel in a sliding window [LST16]. The model was trained and evaluated on 225 prostate biopsy slides randomly selected from a cohort of 238 patients. In order to train patch-level models, the authors collected contour-based annotations from pathologists. Nagpal *et al* proposed a two-stage deep learning system for GG classification of whole slide images from prostatectomy specimens [NFL19, NFT20]. The first stage model was an ensemble deep CNN, which was trained with 112 million labeled patches extracted from 912 slides with pixel-level annotations. In the second stage, a k-nearest-neighbor-based (KNN) model was utilized to aggregate patch-level results.

While these papers demonstrated promising performance in slide-level predictions [NFL19, LST16, NFT20], they required a large number of expensive pixel or patch-level manual annotations for training. Bulten *et al* utilized a semi-automated labeling technique for prostate biopsy slide classification [BPB19, BPB20]. Specifically, the authors used a pre-trained tissue segmentation network to identify tissue areas, within which cancerous regions were localized by a pre-trained tumor detection network. Non-epithelial areas were excluded from identified cancerous regions with an epithelium detection model. Detected epithelial areas from slides with a single Gleason pattern inherited slide-level labels and formed their initial training set for a U-Net model. Slide-level predictions were determined by percentage of Gleason patterns obtained from the segmentation network. However, this framework was built upon three pre-trained preprocessing modules, each of which still required pixel-wise annotations.

2.2.3.1 Multiple instance learning

Multiple instance learning (MIL) is a type of supervised learning, which is proposed to deal with problems with incomplete training labels. Specifically, traditional fully supervised image classification models are usually developed on datasets where each input image has one or more corresponding labels, while MIL framework is proposed to address the challenge that only one label is available for a set of images as shown in Figure 2.5. MIL is firstly proposed to solve the drug activity prediction problem [DLL97]. The drug is considered as effective if it can strongly bind to a target binding site of a molecular, which could have more than one possible three-dimensional shapes, known as conformations. The label about drug effectiveness is only available for a bag of conformations [DLL97, ATH03, Amo13]. Other examples where MIL models can be useful include image segmentation with merely image-level labels [JHE17, XJW17], tumor detection [MGM15, MGM14, QLC16] and classification [HSK16, TBO19, MAM17, WZY19, ITW18, YZP16].

In the MIL, each input data is considered as an instance and a label is available for a bag of instances. The basic assumption of MIL models for binary classification is that the bag is positive, if it has at least one positive instance. This is also referred as the standard multiple instance assumption (SMI). MIL models can be roughly divided into two main categories: 1) bag-level approaches; 2) instance-level approaches.

Instance-level approaches. The discriminative information is considered to be at the instance-level for instance-level methods. Specifically, models are firstly trained to classify instances. Bag-level labels are obtained by aggregating instances scores with different functions such as maximum pooling, mean pooling, noisy-or pooling, noisy-and pooling, convex

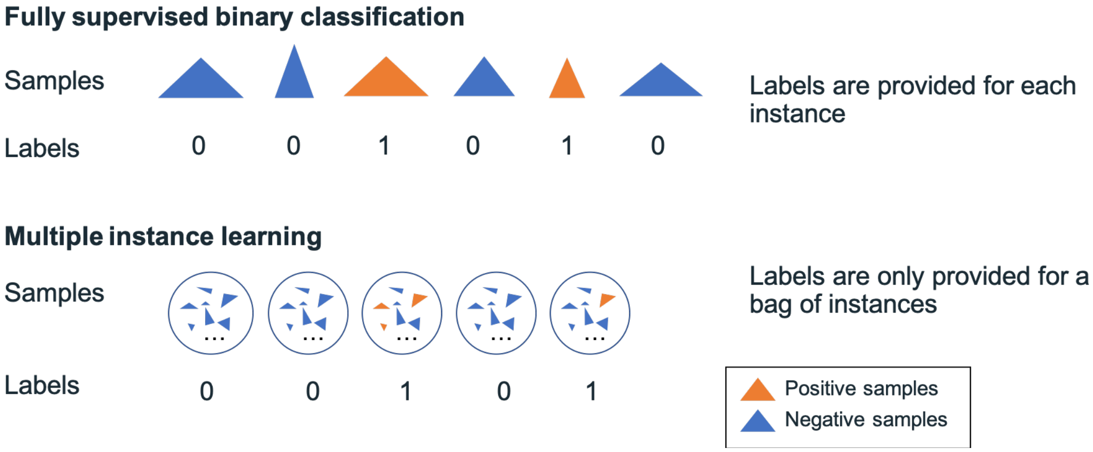
Fully supervised binary classification
Samples
Labels
0
0
1
0
1
0
Labels are provided for each instance
Multiple instance learning
Samples
...
Labels
0
0
1
0
1
Labels are only provided for a bag of instances
Positive samples
Negative samples

Figure 2.5: Differences between fully supervised image classification models and MIL models in a binary classification scenario. Orange triangles denote positive samples and blue triangles represent negative samples. For fully supervised classification tasks, each input data is usually associated with one or more labels as shown in the top row. Yet, MIL models mainly deal with the scenario where a bag of images has one or more labels as demonstrated in the bottom row. Figures are better viewed in color.

maximum operator and *etc* [RD00, ML97, KBF16, FZ17, PC15, ZLV17, ZPV06]. Since labels for individual instance are unavailable for MIL models, instance-level methods usually utilize bag-level classification accuracy to represent instance-level performances.

Models fall in this category include axis-parallel rectangle (APR) [DLL97], diverse density (DD) [ML97], EM-DD [ZG02], multiple instance support vector machine (MI-SVM), mi-SVM [ATH02] and *etc* [Amo13]. The mi-SVM, for example, utilized an iterative training strategy similar as the EM method. It imputed labels for instances in positive bags. The SVM model was trained to optimize the decision boundary for each instance, and then utilized to update instance labels. Label assignment and SVM training steps were iterated until convergence [ATH02].

Besides predicting bag-level labels, instance-level methods can also be used for instance classification, since these models produce predictions for individual instance before classifying bags. However, it has been shown analytically and empirically that the accuracy at instance-level and bag-level is not always consistent and depends on various factors such as the number of instances in the bag and the class imbalance [CCG18, VFB16]. Instance-level models usually achieve inferior performances on bag-level classification tasks [CTL15, ITW18, CCG18].

Bag-level approaches. Different from instance-level models, which consider each instance separately and combine instance-level scores to form bag-level predictions, bag-level methods treat a bag of instances as a whole and directly optimize for bag-level predictions. To address the challenge of bag-level optimization, we can measure distances between two bags (*i.e.,* two sets of points in k dimension) with different functions such as earth movers distance (EMD) [ZML07] and Hausdorff distance [WZ00], which can be plugged in distance-based classifiers such as SVM and KNN for bag-level classification. Kernel functions that computes the similarity between bags can also be used.

Some examples of bag-level methods include MI-Graph [ZSL09], citation-KNN [WZ00], EMD-SVM [RTG00], MIMLSVM [ZZ07], MILES [CBW06], BP-MIP [ZZ04] and *etc.* MILES model, for example, projected each bag into an embedding space where the similarity of a bag to different instances can be represented by its' location. 1-norm SVM was utilized to select discriminative features and build the bag-level classifier.

Empirically, bag-based methods usually demonstrate better performance for tasks where global (*i.e.,* bag-level) predictions are more important. Nevertheless, they suffer from a lack of interpretability and cannot be used for instance-level classification, since instance predictions are often unavailable [ITW18].

Attention-based MIL model Different from pre-defined aggregation function used in many instance-level models, Ilse *et al* developed an attention-based MIL model that replaces fixed aggregation method with a parameterized two-layer neural network (*i.e.,* an attention module) to enable more flexible combination of instance-level information. Specifically, given a bag of n instances $X_i, i = \{1, 2, 3, ..., n\}$, feature vectors extracted from instances V were multiplied by weights W produced from the attention module and formed a bag-level representation. The bag-level feature vector was then forwarded into the final classifier. The learned attention weights can visualize the relative contribution of instances for final prediction, thus, improve the interpretability of the model without sacrificing bag-level prediction performances [ITW18]. The model was utilized to identify epithelial and malignant patches within small tiles extracted from WSI for colon cancer and breast cancer datasets, respectively [GBO08, SAT16]. However, they did not address the challenge of classifying much larger and more heterogeneous WSIs and the potential of using attention maps for relevant regions selection wasn't explored in the paper. Moreover, as pointed out in the paper, learned attention maps were sparse (*i.e.,* not only few relevant instances had large attention weights, while others had small values).

2.2.3.2 Multiple instance learning for medical images

The MIL paradigm assumes that only the global label for a set of instances are available, while labels for individual instance are hidden variables. This fits well for WSI analysis, where the global diagnosis is assigned to the entire slide containing billions of pixels and each pixel or tiles that occupy smaller regions on the WSI can be viewed as instances with unknown labels. §2.2.2 reviews several previous works that used FCN and MIL for weakly-supervised image segmentation. This section mainly focuses on MIL for WSI classification and detection.

Sudharshana *et al* compared multiple MIL-based methods for differentiating benign and malignant histopathological images, including APR, DD, MI-SVM, citation-kNN, a non-parametric MIL method and a MIL CNN-based approach [SPS19]. Models were evaluated on a public dataset containing 8000 biopsy images from 82 patients. The paper showed that MIL-based models outperformed models under the single instance classification framework [SPS19]. Tomczak *et al* developed an instance-level MIL model, which combined patch-level prediction scores with permutation invariant operators such as noisy-or [HS13], integrated segmentation and recognition operator [KRL91] and *etc* [TIW17], for the classification of breast cancer biopsy slides. Campanella *et al* employed an instance-based approach to discriminate between malignant and benign prostate WSIs [CHG19, NFT20]. They considered the top k tiles with the highest probabilities from positive slides after applying the CNN model as pseudo positive training samples, which were updated in each training epoch. In the second stage, they investigated aggregation functions to produce a final slide-level prediction. The model achieved promising performances on three different types of large-scale clinical datasets. However, the more difficult problem of Gleason grading was not investigated in the paper.

2.3 Computer aided progression model

Histopathological slides are known to contain rich information about disease prognosis and are essential in disease diagnosis. Current pathology diagnosis workflow relies on grading systems such as Gleason grading system for prostate cancer, which summarizes diverse histological patterns into certain categories. Yet, these methods may suffer from inter- and intrareader variability [AMJ01]. Also studies have demonstrated that patients within the same category may have heterogeneous outcomes [WSM20, KZL10, LIS14]. Thus, there remains a need for developing computer aided tools that could extract quantitative histopathological features from large-scale WSIs to enable more precise risk stratification for patients. Basic models for survival analysis are presented in §2.3.1. §2.3.2 reviewed several related previous work on using pathomic features for survival analysis.

2.3.1 Survival analysis

The survival analysis is to predict the expected duration until one or more events occur. Each subject i in the dataset for survival analysis usually contains three parts (X_i, t, e_i): the event indicator e_i, the covariates X_i (*e.g.,* features extracted from WSIs of a patient) and the time to event t_i. If the event is observed, $e_i = 1$. Otherwise, the observation is censored $e_i = 0$. The survival function $S(t) = Pr(e_i \geq t)$ describes the probability of the event hasn't occurred at t. The hazard function, which measures the risk of events, models the conditional probability that the event will happen within $[t, t+\delta)$ given it hasn't occurred before as defined in Equation (2.1).

$$\lambda(t|X_i) = \lim_{\delta \to 0} \frac{Pt(t \leq T < t + \delta)}{\delta} \tag{2.1}$$

Larger hazard indicates higher risk of events. The standard linear regression method, however, fails to handle right censored observations. The most commonly used method for survival analysis is the Cox proportional hazard model (CPH) [Cox72]. One of the key assumption for CPH is the proportional hazards function assumption. Specifically, the cox model assumes each covariate x_i has a multiplicative relationship with the hazard and the ratio of hazards remains constant over time. The hazard function for the CPH can be defined in Equation (2.2). It measures the hazard at t given the covariate vector X_i for subject i. λ_0 is the baseline hazard function, which denotes the hazard with all covariate equals to

0. $h(X_i) =\ \exp(\beta^T X_i)$ is the risk function, which represents the relationship between the hazard and predictors (*i.e.,* covariates).

$$\lambda(t|X_i) = \lambda_0\ \exp(\beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_j x_{i,j}) \tag{2.2}$$

To estimate parameters β, we can maximize the partial likelihood that the event is observed for subject i at T_i, given a set of subjects $j \in R_{(T_i)}$ where events hasn't occurred as shown in Equation (2.3) [Cox72, KSC16].

$$l(\beta) = \prod_{i:e_i=1} \frac{\exp(\beta^T X_i)}{\sum_{j\in R_{(T_i)}} \exp(\beta^T X_i)} \tag{2.3}$$

We can take log transform of the likelihood function and minimize the negative log partial likelihood as defined in Equation (2.4):

$$l(\beta) = -\sum_{i:e_i=1} (\beta^T X_i - log \sum_{j\in R_{(T_i)}} \exp(\beta^T X_j)) \tag{2.4}$$

Coefficients learned with the CPH method represent effects of predictors on the hazard, which makes the CPH model easy to interpret. However, the CPH model may fail with high dimensional features (*e.g.,* the number of predictors is larger than the number of samples). This could be potentially be overcome by including penalty terms such as $\mathcal{L}_1$ (*i.e.,* LASSO regression) and $\mathcal{L}_2$ penalties (*i.e.,* ridge regression) [SFH11]. The ElasticNet penalty combines both $\mathcal{L}_1$ and $\mathcal{L}_2$ penalties as defined in Equation 2.5. $\gamma \in [0, 1]$ controls the relative strength of $\mathcal{L}_1$ and $\mathcal{L}_2$ penalties.

$$\arg\min_{\beta}\{l(\beta) + \alpha(\gamma \sum_j |\beta_j| + \frac{1-\gamma}{2} \sum_j \beta_j^2)\} \tag{2.5}$$

The CPH assumes a linear association between predictors and hazard. To address the challenge of modeling non-linear interactions between predictors and the hazard, Faraggi and Simon developed an approach that utilized feed forward neural network to model non-linear relationships [FS95]. They optimized the similar partial likelihood as the CPH, but replaced the linear function $h(X)$ with a neural network. Kartzman *et.at.,* [KSC16] proposed the DeepSurv, which extended the model [FS95] by adding hidden layers and including modern neural network training and regularization techniques such as weight decay, ReLU [NH10], batch normalization [IS15], dropout [SHK14], gradient clipping [PMB12] and *etc.* The DeepSurv model outperformed most state-of-the-art survival models and was able to model complex relationships between covariates and the survival.

2.3.2 Progression-free survival analysis with pathomics features

Many previous work on progression prediction with histopathological images mainly focused on extracting pre-defined morphological features of glands and nuclei from manually identified ROIs, and then correlated extracted representations with survival [DAH18, YZB16, CLL20, LJE19, LRW18].

Chandramouli *et al* developed a model that utilized quantitative histomorphometric features to categorize patients on active surveillance into low- and high-risk group for disease progression [CLL20]. They first segmented nuclei with watershed-based segmentation method. Then 219 handcrafted morphology features were extracted based on segmentation results, including graph-related features, nuclear shape features, nuclear disorder features and cell cluster graph features [CLL20]. In the work by Leo [LJE19], glandular features such as gland shape, arrangement and disorder were used in a Cox model to predict the risk of biochemical recurrence after radical prostatectomy. Lu *et al* extracted shape and orientation features of nuclei from tissue microarray images to predict survival of patients with early-stage breast cancer [LRW18].

Instead of relying on hand-crafted features from pre-defined ROIs, the framework developed by Zhu *et al* utilized adaptive sampling to randomly sample patches from WSIs, which were then clustered into groups. Then they assumed each patch inherited the same survival label from that patient, and trained a patch-wise survival model with the Cox loss function [KSC16, ZYZ17]. Features from clusters with sufficient accuracy were aggregated via the pre-defined function for case-level survival analysis. Ren *et al* considered disease free time as the time variable and proposed a Cox survival model, which combined image features from CNNs and genomic pathway scores, to better predict risk of progression for patients diagnosed with Gleason score 7 prostate cancer [RKG18]. Wulczyn1 *et al* proposed a deep learning system to predict disease specific survival for colorectal patients [WSM20]. The system mainly contained two models: 1) a fully-supervised tumor segmentation model for detecting ROIs; 2) a prognostic model, which extracted features from patches sampled from tumor regions identified by the first model, for survival prediction. The simple mean pooling was used to aggregate patch-level predictions into case-level results [WSM20].

However, no previous study has done to investigate models that can leverage effectiveness of self-supervised deep features, attention-weighted aggregation and spatial distribution of learned features for progression-free survival prediction for prostate cancer. In this dissertation, a deep learning system is developed to predict progression-free survival as recommended in [LLH18]. The progression included biochemical recurrence, distant metastasis, locoregional recurrence and new primary tumor [LLH18].

CHAPTER 3

Semantic Image Segmentation with Multi-scale Information: A Multi-scale U-Net Model

3.1 Overview

Semantic image segmentation is often considered as an important pre-step for quantitative pathological image feature extraction. Previous work have demonstrated promising results on classifying small image tiles with homogeneous tissue contents, which were usually pre-selected by pathologists [FSJ07, NSJ12, GVG13, DFT10]. However, it can be difficult to extend these models for large whole slide image analysis due to the need of selecting and accurately localizing such small homogeneous tiles. This chapter demonstrates a novel multi-scale U-Net model for semantic segmentation of histopathological images with heterogeneous tissue contents from radical prostatectomies. The model can potentially be used to facilitate the estimation of percentage of different Gleason patterns, which is required in determining primary and secondary patterns in prostate pathology diagnosis.

The proposed method utilizes multi-scale information to generate pixel-wise predictions for four tissue classes (*i.e.*, stroma, benign, Gleason 3 and Gleason 4 glands). The dataset used for this study is described in §3.2. Details about the model are described in §3.3. Evaluation and results are summarized in §3.4. In §3.5 we discuss strengths and limitations of the proposed method. This chapter is based on the content of [LSH17] and [IML18].

3.2 Dataset

Our model is developed on the dataset containing histopathological images of prostate radical prostatectomy. Specifically, radical prostatectomy specimens from 20 patients with a diagnosis of Gleason 3 (G3) or Gleason 4 (G4) prostate cancer according to the contemporary grading criteria [FAB12, BME13] were retrieved from archives in the Pathology Department at Cedars-Sinai Medical Center (IRB approval no. Pro00029960) [GIM15]. The specimens were previously stained with hematoxylin and eosin (H & E) for histological evaluation of the tumor.

Slides were digitized by a high resolution whole slide scanner SCN400F (Leica Biosystems, Buffalo Grove, IL). The scanning objective was set to 20x. The output was a color RGB image with the pixel size of $0.5\mu m \times 0.5\mu m$ and 8 bit intensity depth for each color channel. Areas with tumor previously identified by the pathologist were extracted from whole slide images (WSIs) and then saved as 1200×1200 pixel tiles for analysis. 224 tiles were selected by three collaborating pathologists who identified stroma (ST), benign glands (BN), G3 cancer, and G4 cancer containing cribriform and non-cribriform growth patterns. Individual glands and stroma in each tile were annotated manually using a custom graphical user interface [GIM15]. All annotated image tiles were cross-evaluated by the pathologists, and corrections made if there was no consensus. This collection contains: BN (n=32), G3 (n=24), G4 (n=22), G3 and BN (n=29), G4 and BN (n=6), G3 and G4 (n=80), and G3 and G4 and BN (n=31) image tiles. All tiles were normalized to account for stain variability [RAG01].

3.3 Method

3.3.1 Multi-scale U-Net

CNNs achieved promising results on image classification. Unlike classification that classifies the whole image into categories, segmentation requires local predictions for each pixel. To extend the CNN for segmentation tasks, we can apply the model in a sliding window way and produce pixel-wise predictions. However, this method requires extracting patches and applying the model around every pixel, therefore can be inefficient even for images with moderate size.

The FCN proposed by Shelhamer and Long *et al* [LSD15, CPK17] uses up-sampling and fully convolutional layers to generate pixel-wise predictions efficiently in a single pass. The pooling operation makes CNNs relatively invariant to spatial transformations and also reduces spatial resolution of feature maps. To enable making local predictions with global context, the U-Net [RFB15] extends an FCN with a U-shape architecture, which allows features from shallower layers to combine with those from deeper layers [LSD15, RFB15].

One intuitive way of performing semantic segmentation with FCN is to use the entire image as the input. However, training FCN with large images may lead to larger number of parameters. Thus it may require a huge number of samples and also can cause high GPU memory requirement. Downsampling could effectively reduce the image size, but it causes loss of spatial resolution. Some discriminative features such as nuclear features may only be available at higher resolution. To solve these problems, large images are divided into several relatively smaller patches, and the overlap tile strategy is used for seamless segmentation [RFB15]. This method, however, requires the size of the patch to be carefully chosen so that the patch can be segmented with sufficient contextual information. Yet, the size of cellular structures such as glands may vary greatly as shown in Figure 3.1. In addition, both local nuclear patterns available at high resolution and global morphological representations of glands such as shape, which require a larger receptive field but can be available at low-resolution, are important in cancer grading and needs to be considered together. For example, the prominence of nucleoli is an important feature in prostate cancer grading, but it should only be considered in the context of glandular structures to avoid over-diagnosis or under-diagnosis. Single scale input with sufficient receptive fields and spatial resolution may require deeper models and large number of training data.

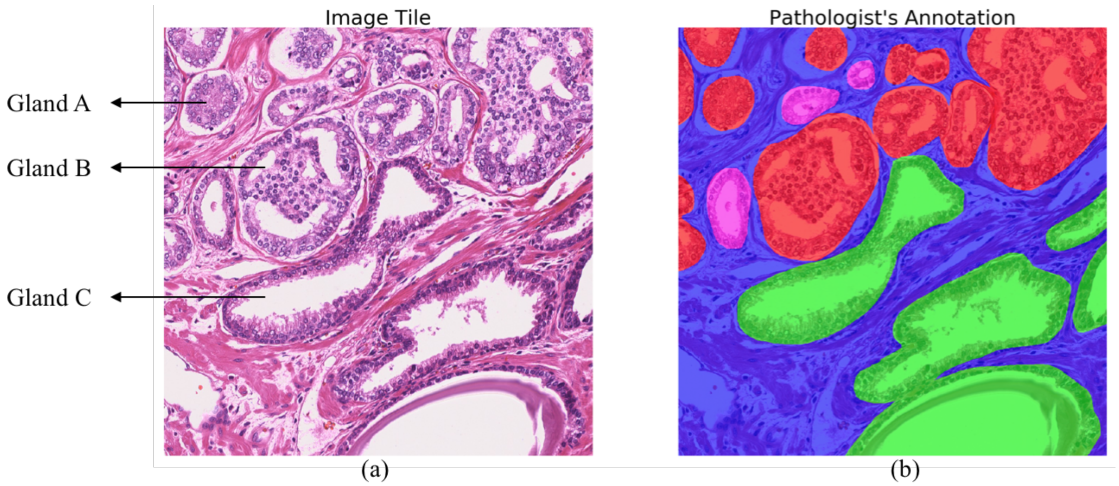

Figure 3.1: Variations in gland size. (a) shows a tile with heterogeneous Gleason grades (G3, G4 and benign glands). Pathologist annotation mask is shown in (b). The high-grade cancer (G4) areas are shown in red, low-grade cancer (G3) areas are denoted as pink, benign glands are indicated by green, and stroma areas are represented by blue. These images demonstrate the heterogeneity of glands both between grades (*e.g.,* glands A and C) and within the same grade (*e.g.,* glands A and B).

The aforementioned challenges motivate the design of a multi-scale architecture that extracts different types of features with inputs of different scales. Specifically, to better segment tissue structures with variable size, we propose a multi-scale U-Net architecture that incorporates patches (sub-tiles) of three different sizes: 400×400, 200×200, and 100×100 to explicitly provide contextual information at multiple scales [HDW17]. The larger patch is designed to capture glandular features, while the smaller patch is designed to extract fine-grained nuclei features. To handle border patches that cause one of these patch sizes to extend past the boundary of a given image tile, the tile is padded with reflection of the border [RFB15]. A detailed overview of our multi-scale U-Net architecture is shown in Figure 3.2. Instead of taking the whole 1200×1200 image tile as input, we divided images into 100×100 subtiles and extracted the three patches of varying size around each of these image subtiles. Features from different sizes of patches were then concatenated together and used as inputs for the multi-scale U-Net model. The commonly used fully connected layer was replaced by a $4 \times 1 \times 1$ convolutional layer that output pixel-wise probabilities for four classes (G3, G3, ST, and BN).

In this experiment, we trained two FCN models. The first was the baseline U-Net model that followed an existing work [RFB15]. The other is the multi-scale U-Net. Both models were trained with batch gradient descent (batch size: 25) and backpropagation. A momentum of 0.9 and a learning rate of 0.05 were used. A heuristic was followed to improve the learning of deep neural network model [KSH17], where the learning rate was decreased by 10x when validation errors stopped decreasing. Models were implemented in Torch7 [CKF11], and the training was done on two NVIDIA Titan X GPUs. The dataset of 20 patients was divided into 10 folds resulting in two patients in each fold. This patient-based cross validation ensured independence of training and testing data.

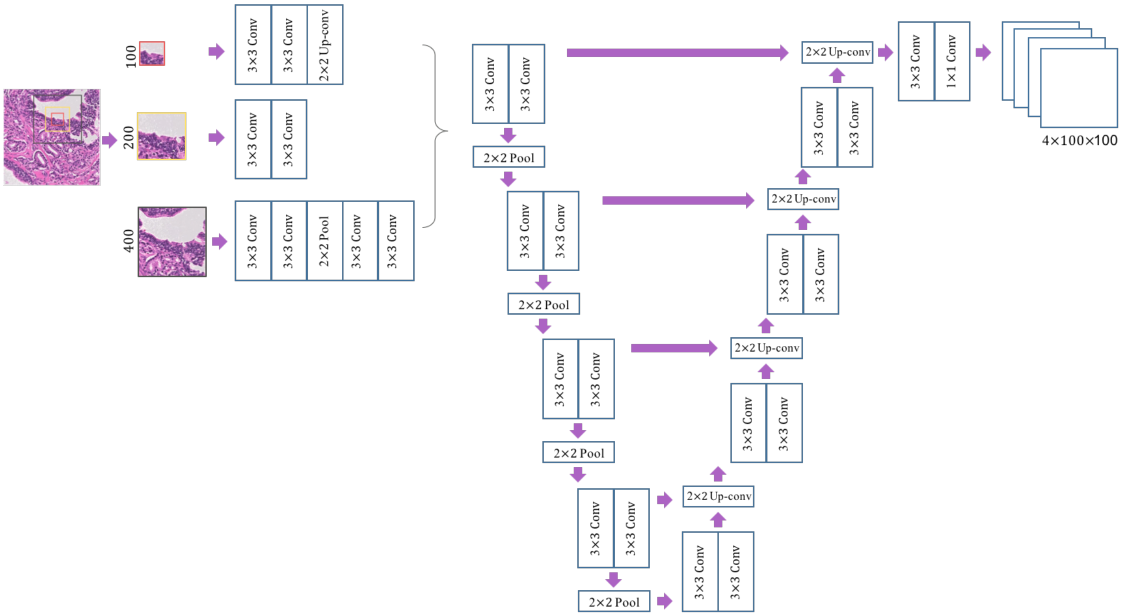

Figure 3.2: Architecture of the multi-scale patch-based U-Net. The whole image was divided into multiple non-overlapping 100 × 100 sub-tiles. To capture contextual information, a 200 × 200 patch (framed in yellow) and a 400 × 400 patch (framed in black) were extracted around each centered 100 × 100 patch (framed in red). Features of different sizes were either down-sampled or up-sampled to 200 × 200, and concatenated into 64 × 200 × 200 feature maps that were input to a U-Net model. The final layer output a 4 × 100 × 100 probability map, each channel of which corresponded to a probability map of one class.

3.3.1.1 Semantic image segmentation with a deep convolutional neural network

For baseline comparison, a deep CNN model was trained to produce pixel-wise class predictions in a sliding window way. The tile dataset was split into a training set containing 187 tiles and a testing set containing 37 tiles. In order to avoid correlations between data in the training and tests sets, tiles belonging to the same patient were restricted to either the training set or the testing set, yielding 17 unique patients in the training set, and 3 unique patients in the testing set (cross-validation was not used due to the large time requirements for evaluating the model).

Specifically, the Inception V3 CNN model [SLJ15] was used with an input size of 299×299 pixels. Patches of size 299×299 were extracted from the pathology image tiles and then used for training and evaluation of the network. The label for any given patch was set to be the true label of the central pixel of the patch. Because there are a large number of possible 299×299 patches (each tile has over 800,000 possible 299×299 patches), it is impractical to train a network on every patch that exists in the dataset. Instead, patches were sampled (with replacement) from the training set using balanced random sampling. In this approach, patches were sampled with equal probability for each class. Within a given class, every potential patch that would fall into the class had equal probability of being sampled. Because of the class imbalance of the dataset, in this methodology, individual potential patches from different classes would have unequal probability of being sampled. Training was performed using an RMSProp (LR $= 0.001$, $\rho = 0.9$, $\varepsilon = 10^{-8}$) [TH17] optimizer using Keras [Cho18] with Tensorflow [AAB16] on two NVIDIA Titan X GPUs with synchronous gradient updates and a batch size per GPU of 50 patches. In order to saturate the GPUs during training, patch sampling was run in threads with separate state; one sampler thread was used per GPU. Training was performed over 25 “epochs” of 100,000 patches.

For evaluation, every possible patch was extracted from tiles in the testing set, and any patches that would have extended outside of the bounds of the original tile were discarded.

Class predictions were obtained for these patches from the network, and each pixel was assigned a class based on the maximum prediction probability for that pixel.

3.4 Experiment and results

3.4.1 Evaluation metrics

Overall pixel accuracy, mean accuracy for each class, and Jaccard index are three commonly used evaluation metrics for multi-class semantic image segmentation. Overall pixel accuracy measures the proportion of correctly classified pixels, however, it can be heavily biased by imbalanced datasets. Mean single-class accuracy calculates the average proportion of correctly classified pixels in each class, which can also be biased by imbalanced datasets and overestimates the true accuracy due to combining multiple negative classes into one inference class [CLP13, EVW10, EEV15]. Jaccard index (J), also known as intersection-over-union, overcomes the limitations of overall pixel accuracy and mean accuracy since it considers both false positives and negatives.

Here, we report Jaccard index and overall pixel-wise accuracy for our models, which can be obtained from a pixel-wise confusion matrix C. C_{ij} is the number of pixels labeled as i and predicted as j. The total number of pixels with label i is denoted as $T_i = \sum_{j=1}^{n} C_{i,j}$, where n is the number of classes. The number of pixels predicted as j is represented as $P_j = \sum_i C_{ij}$ [CLP13]. The Jaccard index for class i is then defined as follows:

$$J_i = \frac{C_{ii}}{T_i + P_j - C_{ii}} \tag{3.1}$$

The overall pixel-wise accuracy (OP) is defined as:

$$OP = \frac{\sum_i C_{ii}}{\sum_i \sum_j C_{ij}} \tag{3.2}$$

3.4.2 Results and discussion

For the pixel-wise deep CNN model, class predictions were produced for a testing set comprising 30,170,133 pixels in 37 tiles across 3 patients. For the standard and multi-scale U-Net models, pixel-wise confusion matrices were summed across all 10 folds. In the first evaluation, true positive, true negative, false positive, and false negative rates for each class were calculated for all pixels in the dataset. Gleason 3 and Gleason 4 predictions were summed into a single inference class (PCa) for evaluation. For comparison, results from a baseline SVM + RF model by Gertych *et al* [GIM15] are also included. The Jaccard index and overall pixel accuracy of each model are reported in Table 3.1. The analysis was also performed without combining Gleason 3 and Gleason 4 into a single class, with performance shown in 3.2. In both cases, the same network (trained on separate classes) was used for prediction.

The multi-scale U-Net architecture achieved the highest Jaccard index in both segmentation tasks: mean $J = 75.5\%$ for 3 class segmentation and mean $J = 65.8\%$ for 4 class segmentation. Both the U-Net and multi-scale U-Net models outperformed the pixel-wise CNN and the SVM-RF model by Gertych *et.al.* [GIM15].

The multi-scale model obtained 2% higher J for segmentation of benign regions. Similar performances were observed when combining G3 and G4 into one cancer class (*i.e.*, U-net got a J of 74.3 %, which was only 0.4% lower than the multi-scale U-Net performance.) However, if G3 and G4 were not combined into one class, multi-scale U-Net achieved around 4% and 1% higher J for G3 and G4 pattern respectively. The relative percentage of G3 and G4 patterns are important in diagnosis and treatment planning (*e.g.,* patients with G4+G3 prostate cancer may have worse prognosis comparing those with G3+G4) [HKZ14]

To evaluation the efficiency of models, the approximated inference time for each model is also measured. It took about 2 hours for pixel-wise deep CNN model, around 3 seconds for U-Net model, and 9 seconds for multi-scale U-Net to generate predictions for a 1200 × 1200 tile on one NVIDIA Titan X GPU. Dense predictions can be much more efficiently produced

Table 3.1: Model performances on segmenting prostate cancer (PCa), benign glands (BN) and stroma (ST).

	$J_{PCa}(\%)$	$J_{BN}(\%)$	$J_{ST}(\%)$	$J_{Mean}(\%)$	$OP(\%)$
Gertych *et al* [GIM15]	49.5	35.2	59.5	48.1	n/a
Pixel-wise CNN	66.0	59.0	71.0	65.0	63.9
U-Net [RFB15]	74.3	70.6	**80.1**	75.0	86.6
Multi-scale U-Net	**74.7**	**72.6**	79.3	**75.5**	**86.7**

Table 3.2: Model performances on segmenting Gleason 4 (G4), Gleason 3 (G3), benign glands (BN), and stroma (ST).

	$J_{G3}(\%)$	$J_{G4}(\%)$	$J_{BN}(\%)$	$J_{ST}(\%)$	$J_{Mean}(\%)$
Gertych *et al* [GIM15][1]	n/a	n/a	35.2	59.5	47.4
Pixel-wise CNN	23.0	25.0	59.0	71.0	45.0
U-Net [RFB15]	45.8	60.9	70.6	**80.1**	64.4
Multi-scale U-Net	**49.8**	**61.5**	**72.6**	79.3	**65.8**

[1] The previous model (SVM+RF) by Gertych *et al* [GIM15]. only addressed three class segmentation by combining G3 and G4 to PCa.

by FCNs.

Segmentation results generated by U-Net and multi-scale U-Net for two representative image tiles are shown in 3.3. Our models performed well in segmenting different tissue types on image tiles with heterogeneous content, but both models struggled with some border areas due to a lack of contextual information. The small high-grade gland marked by a white arrow in the second row in Figure 3.3, for example, was segmented as low-grade gland

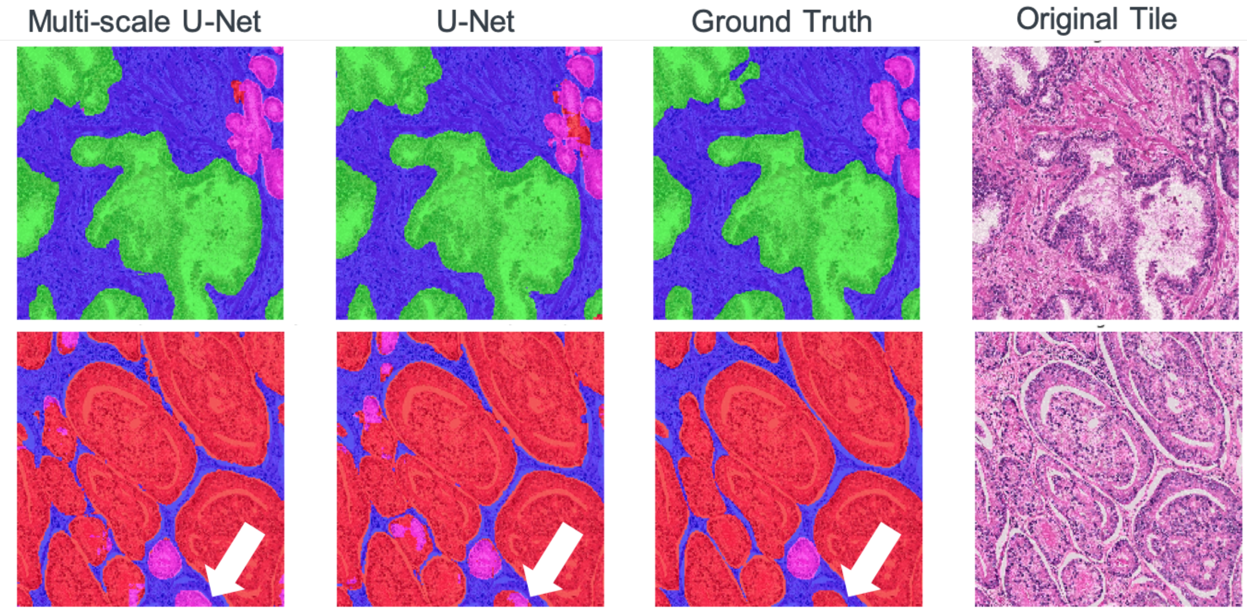

Figure 3.3: Segmentation masks generated by the U-Net and the multi-scale U-Net. Both ground truth masks and predictions are overlaid on original image tiles for easy interpretation. Colors follow the same schema illustrated in Figure 3.1. The first row shows segmentation results for an image tile with three tissue types (benign, stroma, and G3 cancer). The second row shows a representative image tile with three tissue types (stroma, G3 and G4 cancer). White arrows point to border areas that both models struggle with.

by both models.

In cases where global information may be more important for class prediction, the multi-scale U-Net showed superior performance. As shown in Figure 3.4, the single input U-Net misclassified areas with dense nuclei on a large benign gland. However, the multi-scale U-Net was able to segment this area correctly. Though both models could segment large irregular high-grade glands very well as seen in Figure 3.3, they had limited power in segmenting poorly-formed high-grade areas, as shown in the first row of Figure 3.5. Models could detect the approximate location of high-grade cancer, but failed to segment the exact areas. Segmentation performance of both models decreased on tiles with a mixture of small high-

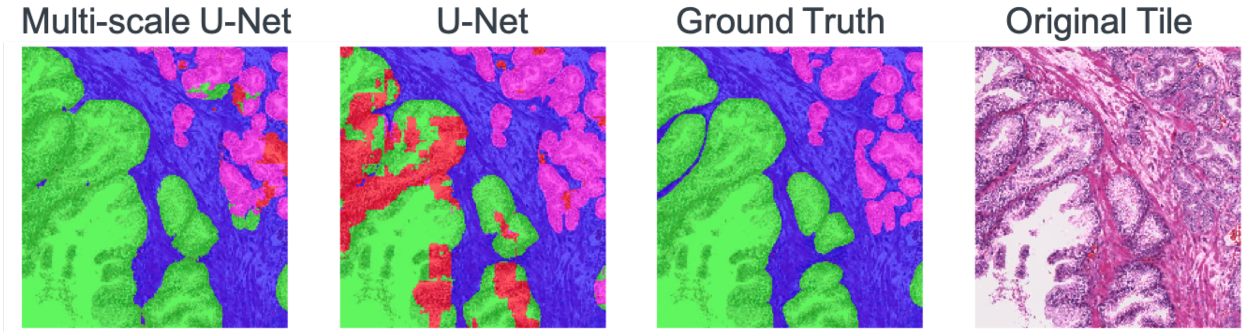

Figure 3.4: Segmentation results comparison for the multi-scale U-Net and the U-Net. Colors follow the same schema illustrated in Figure 3.1. The multi-scale U-Net successfully segmented the large irregular benign gland, while the U-Net with single scale input did not.

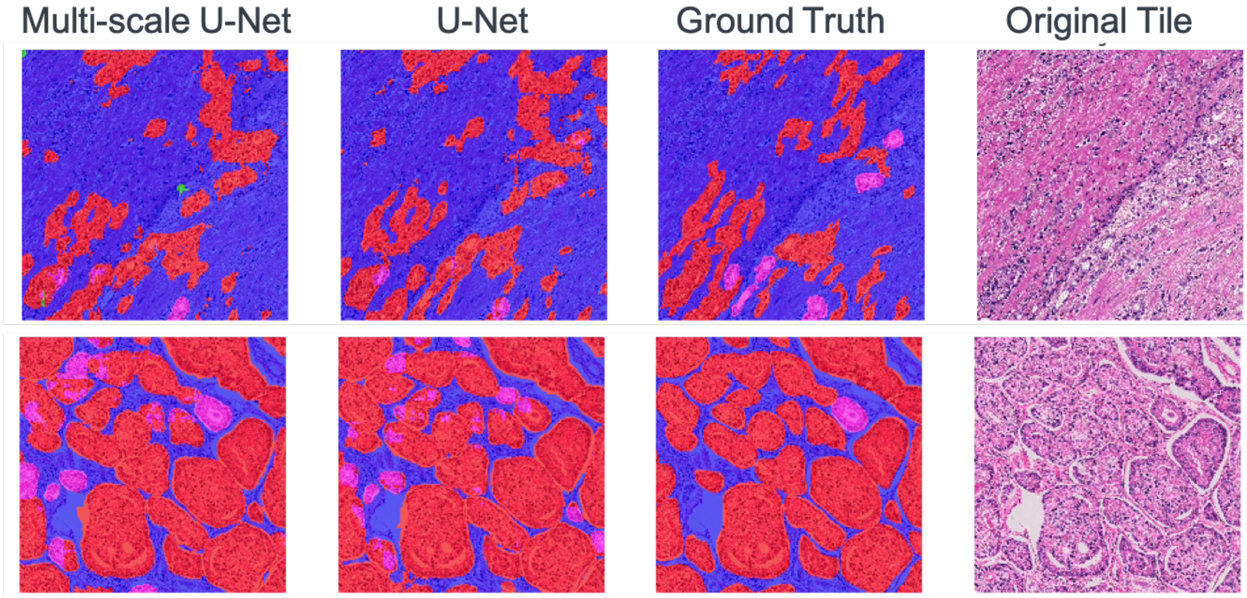

Figure 3.5: Segmentation results for two challenging tiles. Colors follow the same schema illustrated in Figure 3.1. The first row shows an image tile containing G4 cancer with poorly-formed glands. Glands were less differentiated on that tile, likely increasing segmentation difficulty. The second row presents a tile with a mixture of small high-grade glands and small low-grade glands.

grade glands and small low-grade glands. The highest Jaccard indices for G3 and G4 achieved by the multi-scale U-Net were 49.8% and 61.5%, respectively. This reflects the reality that differentiating G3 and G4 is a challenging task, even for pathologists. The inter-observer agreement of clinical pathologists for distinguishing G3 from G4 is between 25% to 47% [AMJ01, GIM15]. A larger training dataset that represents more of the natural variance of these cancer grades could allow for improving the models' ability to discriminate between these classes.

3.5 Summary

In this work, we address the challenge of segmenting different tissue types on heterogeneous histopathological image tiles by using deep learning techniques. The proposed multi-scale U-Net with three types of inputs (400×400, 200×200, 100×100) shows superior performance as compared with the original U-Net. Performances of three different deep learning models (pixel-wise CNN, U-Net, multi-scale U-Net) are evaluated and compared using the Jaccard index and overall pixel accuracy. All three models outperform a reference algorithm on three-class (ST, BN, PCa) segmentation. Both the U-Net and multi-scale U-Net models achieve a higher Jaccard index and require much less inference time than the pixel-wise model. The proposed model is able to explicitly extract discriminative features from different levels and make use of more global information without overly increasing memory requirements during model training.

Though our model achieve promising results on segmenting prostate histopathological images, there are some limitations in this work. Models are only evaluated with patient-level cross validation. Performances on external validation sets could be investigated in the future work. Models are only trained on image tiles, rather than whole histopathological images (*i.e.,* whole slide images). Though our method can be extended to whole image segmentation by splitting these images into non-overlapping tiles, the prediction accuracy for boundary patches could be influenced by lack of contextual information and changes in class balance. The processing time may increase linearly with the size of the whole slide images. Also, our model doesn't perform as well in segmenting G4 cancer with less differentiated glands. Exploring other approaches, such as the use of two separated models with two scales of inputs [WMR16], could improve performance in the future. In addition, our model relies on pixel-wise annotations, which is time-consuming and expensive to obtain. Thus, in Chapter 4 we further extend the model with an EM-based framework, which is able to utilize image-level labels to improve performances of segmentation models.

We also plan to investigate the influence of global versus local features on predicting dense labels, and will perform further evaluations of our models with whole histopathological images and extend our algorithm to a computerized tool which can be used to extract reliable and reproducible quantitative features from histopathological images.

3.6 Appendix

3.6.1 Additional experiments results

We further evaluate the multi-scale model on an extended dataset [IML18] and compare with other FCN models including FCN-8s [LSD15], SegNet-Full [BKC17] and SegNet-Basic [BKC17].

3.6.1.1 Data

Besides 224 tiles (referred as set A) as described in §3.2, 289 newly collected tiles (referred as set B) extracted from radical prostatectomies of 20 patients are included. In addition to G3, BN and ST areas, the new set contains high-grade (G5) cancer regions and G4 with and without cribriform patterns. Slides in set B are scanned by the Aperio scanning system (Aperio ePathology Solutions, Vista, CA). Both sets are scanned at 20x with pixel size of $0.5\mu m \times 0.5\mu m$. Similar as set A, tiles of size 1200×1200 are extracted from slides and then annotated by or under the direct supervision of an expert research pathologist. 5-fold cross validation is used to evaluate model performances.

3.6.1.2 Experiment

FCN-8s, SegNet-Full, SegNet-Basic and multi-scale U-Net model were trained with 10x images downsampled from the original 20x inputs. For FCN-8x, SegNet-Full and SegNet-Basic models, images were generated from 512×512 sub-tiles obtained from the original 1200×1200 tiles and then downsampled to 256×256. For the multi-scale U-Net model, tiles at 10x were divided into 100×100 subtiles, and features were extracted from patches of varying sizes (*i.e.,* 400×400, 200×200 and 100×100) around each of sub-tiles as described in §3.3.

All models were optimized with stochastic gradient descent (SGD) with momentum of 0.9. FCN-8s and SegNet-Full models were initialized with weights pre-trained on PASCAL-

VOC dataset [EVW10], while multi-scale U-Net and SegNet-Basic were trained from scratch. Batch size of 1 and an exponential decaying learning rate started at 1×10^{-4} were used to train the FCN-8x model. SegNet-Basic model was trained with a batch size of 16 and a fixed learning rate of 1×10^{-4}. SegNet-Full was optimized with a batch size of 6 and a fixed learning of 1×10^{-3}. For the multi-scale U-Net, we followed the same heuristic as described in 3.3. FCN-8s, SegNet-Basic and Segnet-Full models were implemented with the Caffe framework [JSD14], while multi-scale U-Net was developed using the Torch7 [CKF11].

We evaluated model performances on segmenting low-grade (LG *i.e.,* G3), high-grade (HG *i.e.,* G4 and G5), benign (BN) and stroma (ST) areas. Same evaluation metrics as described in 3.4: J and OP were used. The average $\bar{J} = \frac{1}{4}\sum_i J_i$ was reported in Table 3.3. J_i is Jaccard index for ith class $i \in HG, LG, BN, ST$.

Table 3.3: Overall model segmentation performances. For each model, results from 5-fold cross validation were gathered into a overall confusion matrix. Evaluation metrics were then computed from the confusion matrix.

Models	#Parameters	OP(%)	$\bar{J}$(%)
FCN-8s	1.3×10^8	87.3	75.9
SegNet-Full	2.9×10^7	82.2	72.1
SegNet-Basic	1.4×10^6	76.2	65.2
Multi-scale U-Net	3.1×10^7	88.5	73.8

Multi-scale U-Net achieved an average OP of 88.5% and an $\bar{J}$ of 73.8%. It was comparable with performance of FCN-8s model, which obtained an average OP of 87.3% and an average $\bar{J}$ of 75.9%. However, the FCN-8s model contained 4 times more parameters than the multi-scale U-Net.

The other observation is that model performances increased as the number of parameters increased with exception of the multi-scale U-Net model. The multi-scale U-Net model achieved around 2% higher $\bar{J}$ and 6% higher OP with similar number of parameters as the SegNet-Full (3.1×10^7 for the multi-scale U-Net and 2.9×10^7 for the SegNet-Full). This further validated the effectiveness of our design of the multi-scale architecture.

3.6.2 Color normalization for whole slide images

Whole slide images are produced by digitizing glass tissue slides that are stained with different types of stains such as hematoxylin (H) and eosin (E) stains. Nuclei are usually stained blue by the hematoxylin, while cytoplasm and connecting tissues are stained pink by eosin. Stains help pathologists to better differentiate histologic structures such as nuclei and glands. However, there could be considerable variations in stains due to different tissue preparation processes, staining protocols, and scanning conditions. Stain variation may affect the performance and the generalizability of machine learning models [CGB17, TLB19]. Stain normalization is one of commonly used approaches to address this challenge.

Reinhard *et al* proposed a general normalization method that converted the source image into LAB color space and then transformed mean and standard deviation of each color channel into targeted values [RAG01]. Different from simple matching of mean and standard deviation, matrix deconvolution-based normalization algorithms leveraged the characteristic of histopathological images, which first found underlying stain vectors and performed color deconvolution to normalize H and E components. For example, Macenko *et al* developed an algorithm that automatically determined stain vectors for color deconvolution [MNM09]. The method assumed each stain has a specific stain vector and the color of each pixel is a linear combination of stain vectors in the optical density space. Singular value decomposition were utilized to estimate stain vectors [MNM09]. This algorithm, however, didn't consider unique structural characteristics information of histopathological images that most tissue regions are only activated by one type of stains. Vahadane *et al* proposed a structure-preserving normalization algorithm that leveraged non-negativity, sparsity, and soft-classification properties of

Table 3.4: Average processing time for different stain normalization methods.

Normalization Method	Average Processing Time (second)
Reinhard [RAG01]	0.3949
Macenko [MNM09]	1.4182
Vahadane [VPS16]	3.5203

histopathological images to estimate stain vectors for stain normalization [VPS16].

We experimented these different normalization methods. Some examples are shown in Figure 3.6 and Figure 3.7. We can see there are significant variations among original images and normalization methods are able to greatly reduce variations. The average time required to run each normalization algorithm on one 1200 × 1200 tile is recorded in Table 3.4. The Reinhard [RAG01] requires the shortest processing time, which is around 10 times faster than the Vahadane [VPS16] method. In this dissertation, the Reinhard [RAG01] normalization method is mostly used for experiments because it's efficient and can achieve relatively good performance.

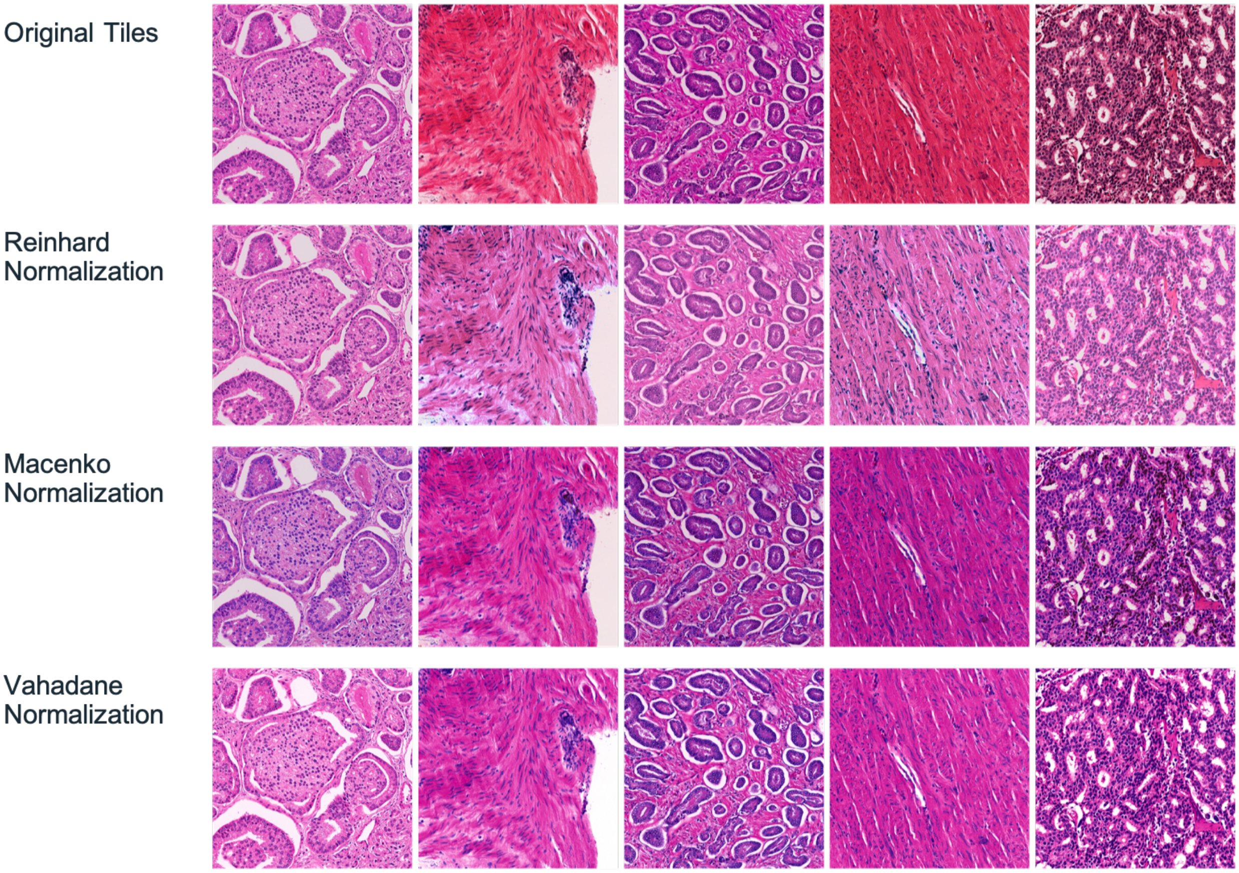

Figure 3.6: Original histopathological images and normalized images with various methods: Reinhard [RAG01], Macenko [MNM09] and Vahadane [VPS16]. Images before normalization contain considerable variations as shown in the first row. Images are better viewed in color.

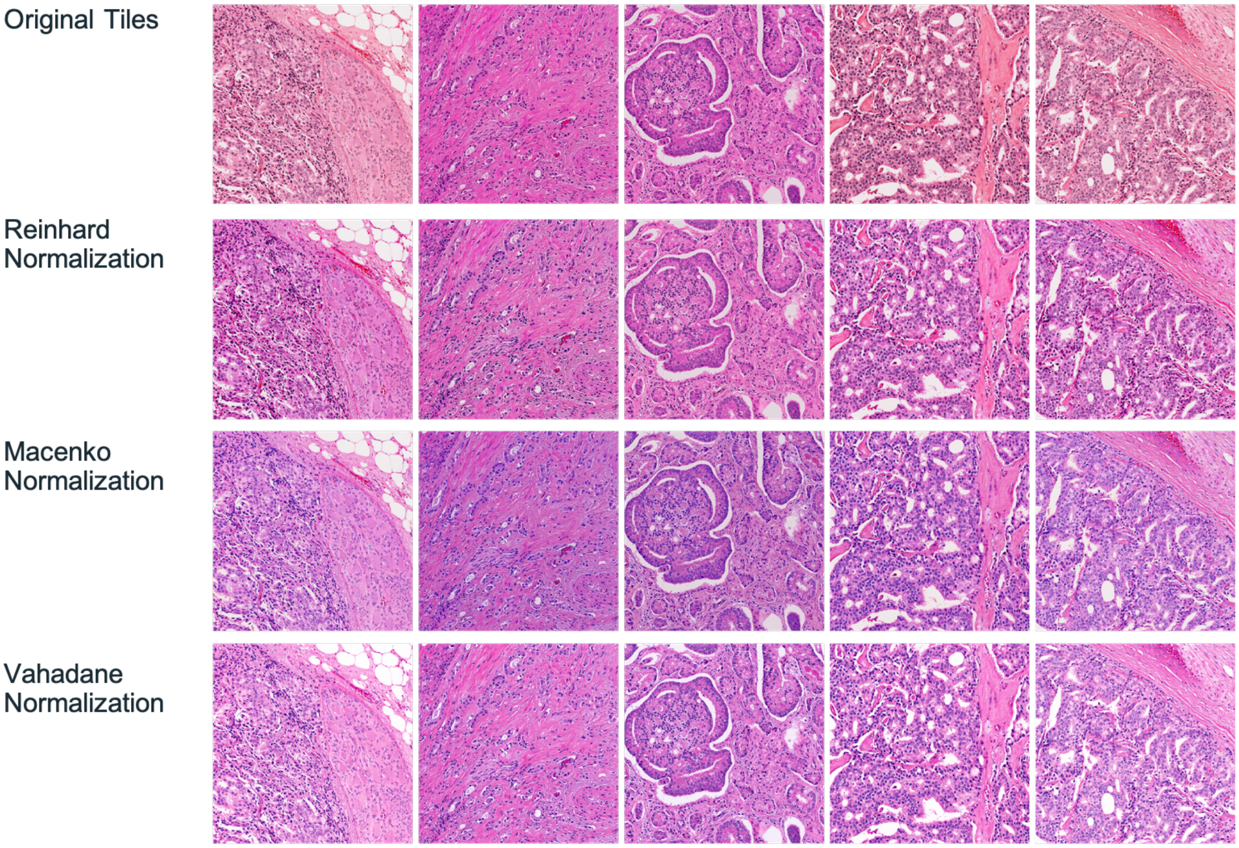

Figure 3.7: Additional examples on original and normalized histopathological images.

CHAPTER 4

Semantic Image Segmentation with Weak Labels: An EM-based Semi-supervised segmentation model

4.1 Overview

A semantic segmentation model would provide Gleason grading for each pixel, which can be used as a preliminary step to extract quantitative pathological image features that are representative of underlying characteristics of tumor. The multi-scale U-Net model presented in the Chapter §3 achieved promising results on segmenting different types of tissues. However, training such a model may require a large-scale dataset to be annotated at gland-level, which would be expensive and time-consuming to produce. In contrast, image-level annotations extracted from low magnification annotations (LMAs) can be generated easily. Figure 4.1 shows the difference between pixel-wise annotations and coarse contour annotations. A model that could utilize image-level labels to train or finetune segmentation models will greatly reduce the cost of collecting annotations and facilitate development of segmentation tools for medical images.

In this chapter we present the EM-based semi-supervised image segmentation model, which can leverage image-level annotations produced from LMAs to augment models trained on the limited data with fine-grained labels. Details on the model are demonstrated in §4.2. In §4.3.1 we introduce the dataset utilized to develop and evaluate EM algorithms. Then we show implementation and training details about the multi-scale U-Net model and EM methods in §4.3.3. In §4.4 we present results of our experiments. Strengths and limitations of this work are discussed in §4.5. Contents of this Chapter are based on [LSH18].

4.2 Method

We use the multi-scale U-Net segmentation model as shown in the Chapter 3. Specifically, k patches of different scales (400×400, 200×200, and 100×100) are extracted around each centered 100×100 patch. Smaller patches are designed to capture high-resolution nuclear features at the center, and larger patches are utilized to extract low-resolution shape features from glands. Deep visual representations from multiple scales are concatenated to generate a semantic segmentation output for the center patch. The detailed architecture of the multi-scale U-Net is shown in Figure 3.2.

To improve segmentation performance, we employ two different types of EM-based models: EM with fixed bias (*EM-fixed*) and EM with adaptive bias (*EM-adaptive*). In weakly supervised segmentation, only image-level labels are available, while pixel-wise annotations are unknown. We denote the label for tile k as $y_k \in Y, k = 1, 2, ..., M$ and the pixel value at location (i, j) as $x_{i,j} \in X$. The label for each pixel is considered as the hidden variable $z_{i,j} \in Z$. The complete data is $\{X, Z\}$.

To maximize the marginal likelihood of observed data as defined in Equation (4.1), the EM algorithm iteratively alternates between making guesses about the hidden pixel labels $z_{i,j}$ in the E-step and finding the optimal model parameters θ that maximize $p(Z|X, \theta)$ in the M-step [Bis06, GC11, PCM15]. Here, we can adopt an FCN-based model (multi-scale U-Net) to produce pixel-wise probability maps $p(Z|X, \theta)$.

$$P(X|\theta) = \sum_{Z} P(X, Z|\theta) \tag{4.1}$$

However, this approach doesn't consider information from image-level labels. Also it may fail because of the singularities of the log-likelihood function [GC11]. For example, the

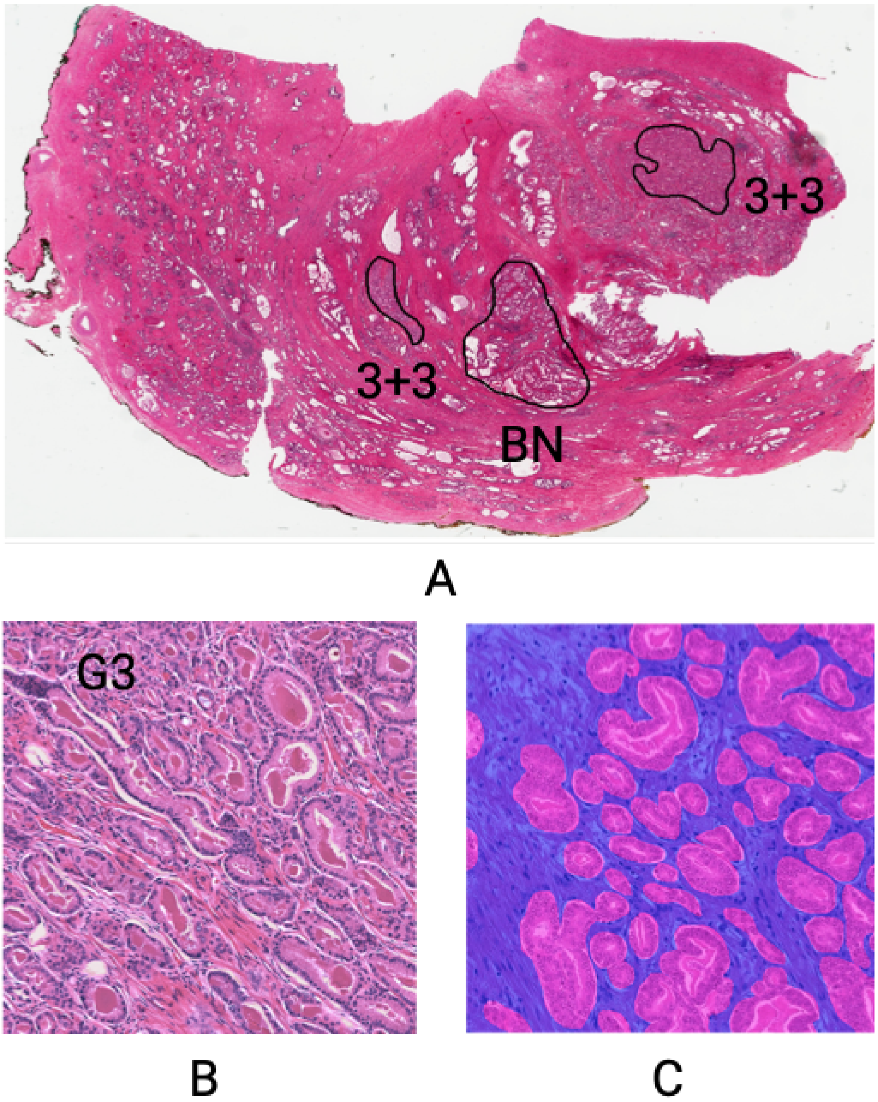

Figure 4.1: Differences between coarse contours and pixel-wise annotations. (A) A whole slide image with contour annotations visualized at 0.4x. Many tiles can be extracted from these contours. (B) A 1200 × 1200 tile sampled from one of the G3+G3 contours on A. It only has an image-level label of G3 inherited from the contour-level label. (C) A tile with pixel-wise annotations. The low-grade cancer (G3) and stromal areas are indicated by pink and blue colors respectively. (Figures are best viewed in color.)

model could converge to a point that predicts most pixels to be stroma. To prevent such degeneracy, we constrain the model output based on the image-level labels. One simple method to incorporate image-level labels is to apply a fixed bias on the output probability maps. Specifically, the probability of any class except the labeled class and stroma is set to 0, and the fixed bias β will be applied to incorporate our belief that the pixel has β probability to be classified as the labeled class. Assuming that the model will output a probability $P(y_i^j)$ for class j of pixel i in a tile labeled as K, and the stromal class is represented as S, the updated probability $P^{'}(y_i^j)$ can be calculated by Equation (4.2):

$$P^{'}(y_i^j) = \frac{1}{T}\begin{cases} 0, & j \neq K \ or \ S \\ \beta P(y_i^j) & j = K \\ (1-\beta)P(y_i^j) & j = S \end{cases} \tag{4.2}$$

where $T = \beta P(y_i^{j=K}) + (1-\beta)P(y_i^{j=S})$. The method encourages pixels to be classified as the tile-labeled class or stroma. Yet, the hyperparameter β could affect how much percent of pixel will be classified as the desired class. Thus β has to be carefully selected to improve performance. Also this method may fail if the percentage of classes varies largely for different images.

To address the challenge of searching for a fixed β for each class, we proposed an adaptive bias to match the distribution of latent pixel-wise labels to the prior distribution $Q(Z)$. The proposed method is based on the assumption that the distribution of epithelial areas versus stroma is similar for tiles within the same cancer grade, but different between grades (*e.g.,* in high-grade tiles, cancerous cells infiltrate into surrounding tissues, which results in reducing of stromal areas). In practice, at each E-step we adaptively select the bias β for each class to be applied on the output probability map $(Z|X, \theta)$ by minimizing the Kullback-Leibler (KL) divergence between the prior distribution and the average distribution derived from model outputs given current parameter θ, and bias β settings. If we denote the total number of available tiles in the fully-supervised training dataset as N, the label for each pixel as y_j and the total number of pixels of each image as M, for Gleason grade g, the $Q(Z^g_{EP})$ can be calculated by the Equation (4.3), and $Q(Z^g_{ST}) = \mathbb{1} - Q(Z^g_{EP})$. $\mathbb{1}$ is the indicator function.

$$Q(Z^g_{EP}) = \frac{\sum_i^N \sum_j^M \mathbb{1}_{(y_{j,j \in EP})}}{N \times M} \tag{4.3}$$

In practice, at each E-step we adaptively select the bias β for each class to be applied on the output probability map $P(Z|X, \theta)$ by minimizing the Kullback-Leibler (KL) divergence between the prior distribution and the average distribution derived from model outputs given current parameter θ, and bias β settings.

The following is an overview of our EM-based approach with adaptive bias:

Initialization: Parameters obtained from the multi-scale U-Net model trained on a small fully-annotated dataset (135 tiles) with pixel-wise annotations are used as the initial point θ^0.

E-step: Calculate $P(Z^t|X, \theta^t)$ based on current parameters θ^t. Generate the average distribution of Gleason grade predictions $H(Z^t)$ from probability maps $P(Z^t|X, \theta^t)$. A class-specific adaptive β is applied on probability maps in order to minimize the KL divergence between $H(Z)$ and the prior distribution, $Q(Z)$. Updated probability maps are calculated by Equation (4.1) with $\beta = \beta^*$:

$$\beta^* = \arg\min_{\beta} \sum_{t \in G} KL(H(Z^t_i), Q(Z_i)), \ G = \{\text{Epithelium, Stroma}\} \tag{4.4}$$

M-step: Update the model parameters. The multi-scale U-Net is trained based on the updated pixel-wise label produced in the previous E-step:

$$\begin{aligned} \theta^{t+1} &= \arg\max_{\theta} Q(\theta, \theta^t), \ \text{where} \\ Q(\theta, \theta^t) &= \sum \text{log} P(Z^t|X, \theta) \end{aligned} \tag{4.5}$$

Iterate E-step and M-step until convergence. To further improve the performance of this EM-based algorithm, we add a small portion of labeled patches from the initial strongly annotated dataset in each batch during SGD training. Figure 4.2 shows an overview of our semi-supervised segmentation approach.

4.3 Experiment

Here, we introduce the dataset utilized to develop and evaluate EM algorithms in §4.3.1. Then we show implementation and training details about the multi-scale U-Net model and EM methods in §4.4.

4.3.1 Dataset and image preprocessing

Our EM-based semi-supervised models were evaluated using a dataset obtained from the Department of Pathology at Cedars-Sinai Medical Center that consists of data from three different cohorts:

(A) 224 tiles with a size of 1200 × 1200, which contain stroma (ST), benign glands (BN), low-grade (G3) and high-grade areas (G4 with cribriform and non-cribriform glands) extracted from slides of prostatectomy specimens of 20 patients [GIM15, LSH17]. These tiles were annotated at pixel-wise level by consensus of three uropathologists [GIM15].

(B) 289 tiles with a size of 1200 × 1200, which contain ST, BN, low-grade, and high-grade (G4 and G5) areas obtained from slides of 20 patients. These tiles were annotated in a similar manner as set (A) by or under direct supervision of an expert research pathologist [IML18].

(C) A research pathologist provided coarse annotations on 30 whole slide images from prostatectomies of 30 patients by circling and grading the major foci of tumor as either low-grade (LG), high-grade (HG) or BN areas using the Aperio ScanScope software [IML18].

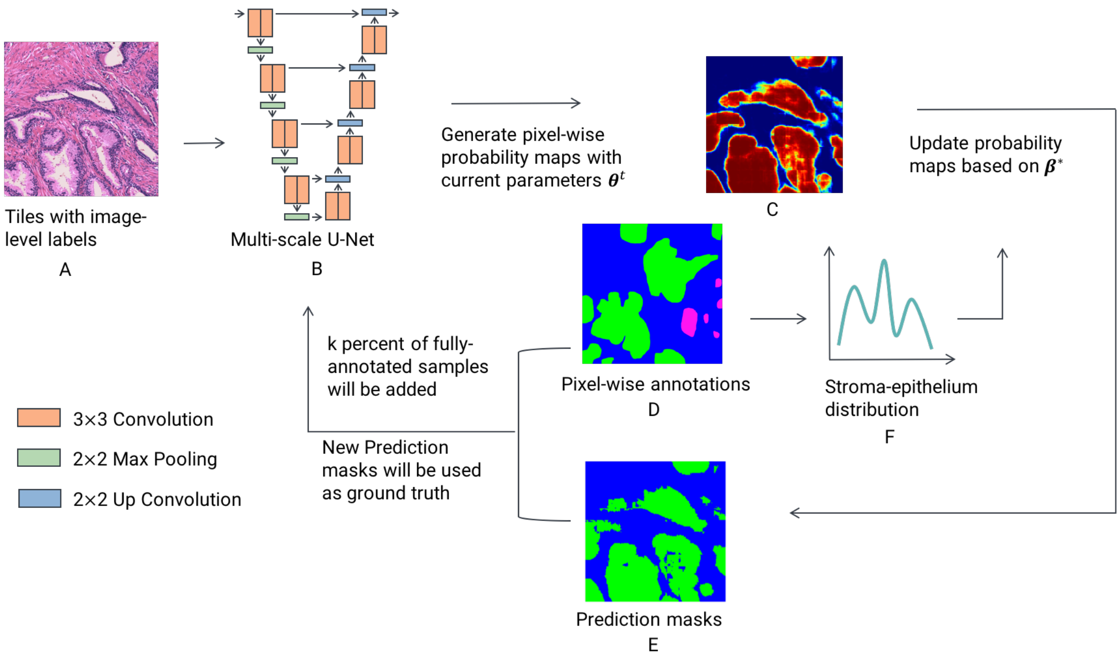
Tiles with image-level labels
A
Multi-scale U-Net
B
Generate pixel-wise probability maps with current parameters $\boldsymbol{\theta}^t$
C
Update probability maps based on $\boldsymbol{\beta}^*$
Pixel-wise annotations
D
Stroma-epithelium distribution
F
k percent of fully-annotated samples will be added
New Prediction masks will be used as ground truth
Prediction masks
E
3×3 Convolution
2×2 Max Pooling
2×2 Up Convolution

Figure 4.2: Overview of EM-based semi-supervised semantic segmentation. LMAs are generated by pathologists so that enclosed regions only contain tissues of the designated label (*e.g.,* A 'G3+3' contour should contain purely G3 glands and stroma, devoid of benign glands or glands of other grades). EM-based algorithms are initialized with a multi-scale U-Net (as shown in B) trained on small amount of tiles with gland-level annotations, and trained on tiles with only image-level labels (as shown in A) extracted from LMAs on histopathological slides. In the E-step, the current model is applied to generate pixel-wise probability maps (as shown in C). To prevent the model from degeneracy, these probability maps are updated by a bias that has been adaptively selected by minimizing the KL divergence between the prior stroma versus epithelium distribution and the average model output distribution. Prediction masks (as shown in D) generated from the E-step are utilized to optimize model parameters in the M-step. To improve training, a small portion of patches with gland-level annotations are combined with patches with image-level labels in each batch. The EM-based method will iteratively update segmentation masks and model parameters until convergence. (Figures are best viewed in color.)

Tiles extracted from the contour were annotated with the same tile-level label as the contour.

The scanning objective for all slides was set to 20x (0.5 μm per pixel). Tiles were normalized using color transfer algorithm [RAG01] to account for stain variability. 60% of the tiles from set (A) were used to train a multi-scale U-Net model and the remaining 40% were used to validate model hyperparameters. EM-based approaches initialized by that supervised model were trained to further improve semantic segmentation performance on around 1,800 weakly labeled tiles extracted from annotated contours in set (C). Hyper-parameters were tuned on 89 left-out tiles from set (A) and model performances were evaluated on set (B) such that tiles from the same patient were not included in both training and testing.

4.3.2 Evaluation metric

We utilized the similar evaluation metrics as in §3.4: overall pixel accuracy (OP) and Jaccard index (J). OP computes the proportion of correctly pixels, which can be easily biased by imbalanced datasets. J, also referred to as the intersection-over-union (IoU), can overcome the class imbalanced problem since it considers both false positives and negatives.

4.3.3 Details on model implementation and training

Given the large number of model parameters, we adopted two typically used regularization strategies: batch normalization (BN) and dropouts. The BN layer was applied after each convolutional layer except the final fully convolutional layer [SLJ15]. Dropout layers with 0.5 probability were added in the deepest stage of the multi-scale U-Net [RFB15, LSH17]. The initial fully supervised multi-scale U-Net was trained on 135 tiles with batch stochastic gradient descent (batch size: 25). EM-based models were initialized with the multi-scale U-Net, and trained with stochastic gradient descent (batch size: 25) in M-steps. Hyper-parameters (*e.g.,* learning rate, number of epochs, weight decay, *etc.*) were tuned on the validation set. The best result was obtained by using a momentum of 0.9, 0.0005 weight decay and a learning rate which was initialized as 0.005, reduced to and fixed at 0.001 after 5 epochs. For EM-fixed models, we used a β value of 0.6 according to the average stroma-epithelium distribution across all classes. In EM-adaptive training, the optimal β for each epithelial class was determined based on the prior distribution. For comparison, we implemented the adaptive method in [PCM15].

Models were implemented in Torch 7 [CKF11] with two NVIDIA Titan X GPUs. Multiple separate data loading threads were used to accelerate training and testing. The average time required to generate a prediction mask for one $1,200 \times 1,200$ tile was around 9 seconds.

4.4 Results

Table 4.1 shows J and OP for models: EM adaptive model w/o fully annotated samples proposed in [PCM15], EM-fixed w/o fully annotated samples (EM-fixed w/o), EM-fixed with 10% fully annotated samples (EM-fixed w 10%), and *etc.* The initial multi-scale U-Net trained with 135 strongly annotated tiles achieved a $\bar{J} = 35.90\%$ on an independent test set. EM-fixed and EM-adaptive models improved segmentation performance by incorporating information embedded in weakly labeled tiles extracted from contours of prostatectomy slides. Using an adaptive threshold resulted in significant improvements in J for low grade glands, high grade glands, and stroma ($p < 10^{-5}$, $p < 10^{-25}$, and $p < 10^{-18}$, respectively), and a non-significant decrease in J for benign glands ($p = 0.18$). The average J and OP both significantly improved when using an adaptive threshold ($p < 10^{-18}$ and $p < 10^{-19}$, respectively). The baseline EM model w/o fully annotated samples [PCM15] achieved $\bar{J} = 42.32\%$ and $OP = 71.84\%$, which was significantly lower than the EM-adaptive w/o ($p < 10^{-14}$ and $p < 10^{-16}$, respectively).

To analyze the contribution of fully annotated samples, different percentages (10%, 30%, 60%, and 90%) of fully annotated patches were mixed with weakly labeled samples in each mini-batch during training. The performance of both EM-fixed and EM-adaptive models improved by adding a small portion of fully annotated samples. For example, the EM-adaptive model with 10% fully annotated samples achieved $\bar{J} = 47.78\%$, which was about 3% higher compared to the EM-adaptive model without fully annotated samples. In addition, the EM-adaptive models consistently performed better than the EM-fixed models. The overall highest $\bar{J}$ was obtained at 49.47% by the EM-adaptive model with 30% fully annotated samples. In Figure 4.3, we show visual comparisons for semantic segmentation on some representative tiles from the test set. Results for the initial fully-supervised multi-scale U-Net model were shown in the first column, while examples for the EM-fixed and the EM-adaptive were presented in the second and the third column respectively.

4.5 Discussion

In this chapter, we demonstrate that EM-based algorithms can learn visual representations from weakly annotated histopathological slides. A small portion of strongly labeled samples increased J on low grade and benign epithelium, but reduced J on high grade and stromal areas, possibly because the limited fully-annotated dataset did not contain all types of high grade tissues (G4 and G5). Moreover, adaptive biases based on prior knowledge of stroma-epithelium distribution lead to better EM training.

As shown in Table 4.1, the initial multi-scale U-Net model only achieved a $\bar{J}$ at 35.90%. The initial model failed to capture the HG glands and erroneously classified those areas into LG, or BN as shown in the first row of Figure 4.3. Since we started with a training set with only 135 fully annotated tiles, the model may have been overfitted to this small dataset and did not generalize well to unseen samples. Furthermore, the HG class in our initial supervised training set only contained non-cribriform G4 and cribriform G4 growth patterns, but not G5 areas with hardly distinguishable glands. We would argue that the multi-scale U-Net may not have sufficient knowledge about visual representations of G5 areas.

Initialized with this undertrained model, our EM-based approaches were able to leverage

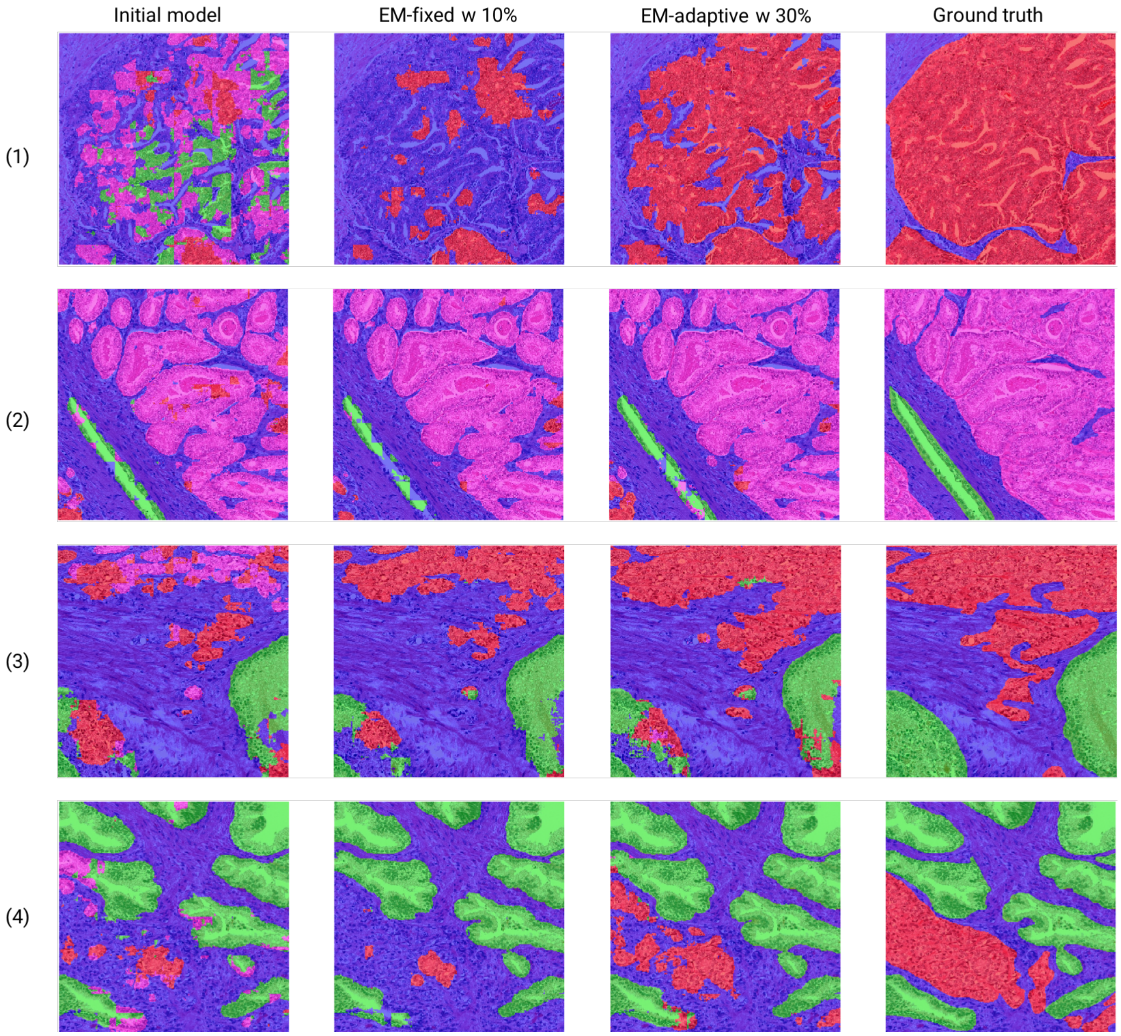

Figure 4.3: Segmentation masks for tiles in the test set. Stromal and benign areas are denoted by blue and green colors respectively. The high-grade (G4, G5) and low-grade (G3) cancer areas are represented by red and pink colors respectively. The first column shows that the initial model delivers inferior performance in segmenting epithelial areas, likely due to the small amount of available supervised training data. Both EM-based models (shown in the second and third columns) are able to improve segmentation performance using weakly labeled tiles. The best performance is achieved by adding 0.3 strongly labeled tiles during EM training. (Figures are best viewed in color.)

Table 4.1: Model performances on segmenting stroma, high-grade (HG), low-grade (LG), and benign (BN) glands.

	$J_{LG}(\%)$	$J_{HG}(\%)$	$J_{BN}(\%)$	$J_{ST}(\%)$	$\bar{J}(\%)$	$OP(\%)$
Multi-scale U-Net [LSH17]	25.80	27.73	24.24	65.83	35.90	64.72
Papandreou *et al* [PCM15]	30.80	50.13*	19.66	68.60	42.32	71.84
EM-fixed w/o strong labels	33.29	44.89*	23.11	67.26	42.14*	**71.11***
EM-fixed w 0.1 strong labels	46.84	30.62	35.93	62.04	**43.86***	67.28*
EM-fixed w 0.3 strong labels	43.59	28.17	35.11	61.76	42.16	66.62*
EM-fixed w 0.6 strong labels	41.82	26.43	30.59	61.23	40.02	65.27
EM-fixed w 0.9 strong labels	39.30	25.61	30.17	60.77	38.96	64.58
EM-adaptive w/o strong labels	33.20	52.01*	23.15	70.27*	44.66*	73.87*
EM-adaptive w 0.1 strong labels	49.67	42.08*	33.45	65.90*	47.78*	71.92*
EM-adaptive w 0.3 strong labels	48.25	49.58*	31.20	68.85*	**49.47***	**74.79***
EM-adaptive w 0.6 strong labels	46.36	42.17*	29.65	66.23*	46.10*	71.61*
EM-adaptive w 0.9 strong labels	40.07	35.62*	31.36	64.10	42.79*	68.65*

* Denotes significant improvement over multi-scale U-Net using Wilcoxon signed-rank tests and Bonferroni correction for multiple comparisons.

rich information embedded in the large-scale weakly annotated dataset. Table 4.1 shows that all EM-based methods outperformed the initial model by a large margin. The $\bar{J}$ improved over 5% by most EM-based models, which demonstrates the ability of our semi-supervised algorithms in extracting useful signals from weakly labeled data. Both fixed and adaptive biases were imposed on pixel-wise probability maps to encourage pixels to be classified as the labeled class or stroma. This significantly reduced the possibility of misclassification of epithelium, such as predicting pixels in an LG tile as HG. As shown in the first and third rows of Figure 4.3, the initial model predicted many HG areas as LG, BN or ST. However, the EM-adaptive model with 0.3 strong labels correctly identified the approximate location of HG tissues, although it might be challenging to get perfect segmentation for HG areas, which have less recognizable glandular boundaries and may infiltrate into surrounding tissues. This kind of imprecise segmentation may be acceptable clinically, since localization of HG areas is considered to be more critical than accurate segmentation.

The baseline EM model [PCM15] adaptively selected bias to constrain for each tile: at least ρ percentage of the tile to be predicted as foreground (epithelium). However, this method didn't take account of differences of stroma-epithelium among individual classes (*e.g.,* High-grade areas tend to have more epithelium). As shown in Table 4.1, the baseline EM model achieved a similar performance as the EM-fixed model.

Different from the baseline EM model, the proposed bias was adaptively selected by minimizing the KL divergence between the model output distribution and the prior stroma-epithelium distribution. These models selected the optimal bias at the cost of longer training time since prediction maps had to be updated whenever a new bias was applied. We found that adding a small percent of strongly labeled data from the initial training set significantly improved model performances. However, adding too much strongly labeled data might prevent the model from learning new information from the large-scale weakly annotated dataset and lead to suboptimal performance. As seen in Table 4.1, the EM-adaptive model with 30% fully annotated samples achieved the highest $\bar{J} = 49.47\%$ and $OP = 74.79\%$.

There are several limitations in our work. First, we assume that each annotated contour contains one type of epithelium (BN, HG or LG), and tiles within the contour inherit its annotation as image-level labels. However, there still may be a very small portion of areas with different Gleason grades. In future work, we plan to extend our current EM-based approach to a multi-class weakly supervised model. Second, we only use a multi-scale U-Net proposed in our previous work as the backbone to generate segmentation masks. In future work, different state-of-the-art multi-scale architectures will also be explored and plugged into our semi-supervised training pipeline. In addition, cross-applicability is important for models that can be extended to computer aided diagnosis tools. We also plan to evaluate our approaches on whole slide images from different institutions, which may have distinct staining or scanning protocols.

4.6 Summary

In this chapter, we present an EM-based semi-supervised model to leverage useful representations embedded in large-scale weakly annotated datasets. Adaptive biases incorporated prior knowledge on stroma versus epithelium distributions and are employed to prevent the model from predicting everything as stroma. The learning of the EM-based models is further improved by combining some fully annotated samples in each mini-batch during training. Our best semi-supervised EM-based approach achieves an $\bar{J}$ of 49.47% on an independent test set, which is 14% higher than the supervised model. The result demonstrates that our semi-supervised model could improve semantic segmentation performance without requiring a very large dataset with time-consuming and costly pixel-wise annotations from pathologists.

CHAPTER 5

Whole Slide Image Classification: A Multi-resolution Multiple Instance Learning Model

5.1 Overview

In Chapter 3 we discuss the segmentation model that can leverage information from both gland-level and nuclei-level. The EM-based framework as described in Chapter 4 further enables the segmentation model to be trained with image-level labels instead of relying on expensive pixel-wise annotations. However, these models only focus on analyzing tiles extracted from large-scale whole slide images. How to effectively identifying regions of interest (ROIs) and combining information from these regions still remains as challenges. This Chapter details a multi-resolution multiple instance learning (MRMIL) model for whole slide image (WSIs) classification and ROI detection.

Different from most existing studies, which rely on highly curated datasets with fine-grained manual annotations at pixel- or region-level, the MRMIL model can be trained with only slide-level labels obtained from pathology reports. Similar to how WSIs are typically reviewed by pathologists, the proposed model scans through the entire slide to localize suspicious regions at a lower resolution (*i.e.*, at 5x), and then zooms in on the suspicious regions to make grade predictions (*i.e.*, at 10x). The model can potentially be utilized as a second reader for Gleason grading, and the produced attention map can be used to help pathologists quickly localize suspicious areas. More details on the model design are demonstrated in §5.2. We also perform experiments to compare the proposed model and other related works, as shown in §5.3. Results of comparison experiments are presented in §5.4. In §5.5 and §5.6 we conclude the chapter by summarizing strengths and limitations of this work. Contents of this Chapter are based on the [LLG19, LLS20].

5.2 Method

5.2.1 Problem definition

Due to the enormous size of WSIs, slides are usually divided into smaller tiles for analysis. However, different from works that utilized fine-grained manual annotations, our model is developed on the dataest with only slide-level labels (*i.e.,* We don't have labels for each tile, instead, we only have a slide-level label for a set of tiles.). Therefore, we formulate the WSI classification problem in the MIL framework. Specifically, a slide is considered as one bag. k tiles of size $N \times N$ extracted from the bag are denoted as instances within the bag, each of which may have different instance-level labels $y_i, i \in [1, k]$. During training, only the label for a set of instances (*i.e.*, bag-level) Y is available. Based on the MIL assumption, a positive bag should contain at least one positive instance, while a negative bag contains all negative instances [ATH03, DLL97, Amo13, CBP19] in a binary classification scenario, as defined in Equation (5.1).

$$Y = \begin{cases} 0 & \text{iff } \forall i \in [1, k], y_i = 0 \\ 1 & \text{otherwise} \end{cases} \tag{5.1}$$

We build our system upon a bag-level MIL model, which combines instance-level representations into a bag-level feature vector for classification using an aggregation method. Instead of relying on a pre-defined function, such as maximum or mean pooling [ITW18], our model utilizes a parameterized attention module that aggregates instance features and forms the bag-level representation. Figure 5.1. shows the overview of our model.

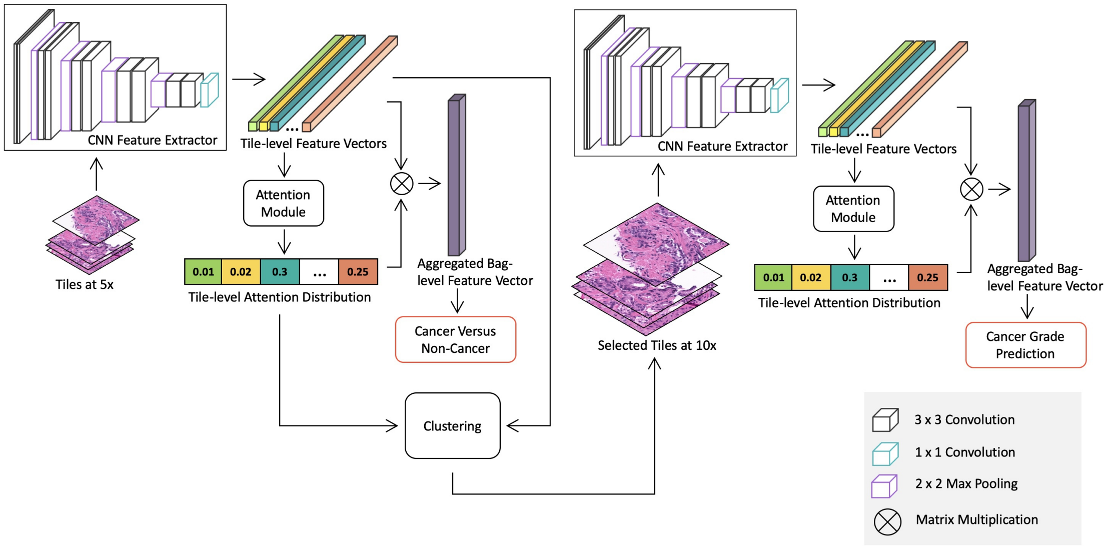

Figure 5.1: Overview of the proposed whole slide image detection and classification model. The model consists of two stages: a cancer detection stage at a low magnification and a cancer classification stage at a higher magnification for suspicious regions. Both stages contain a CNN feature extractor, which is trained in the MIL framework with slide-level labels. Specifically, the detection stage model is trained with all tiles extracted from slides at 5x to differentiate between benign and malignant slides. The attention module in the detection stage model produces a saliency map, which represents relative importance of each tile for predicting slide-level labels. Then we use the K-means clustering method to group tiles into clusters based on tile-level features. The number of tiles selected from each cluster is determined by the mean of cluster attention values. Discriminative tiles identified by the detection stage model are then extracted at 10x and fed into the classification stage model for cancer grade classification.

5.2.2 Attention-based MIL with instance dropout

In the attention-based MIL model, a CNN is utilized to transform each instance into a d dimensional feature vector $\mathbf{v}_i \in \mathbb{R}^d$. A permutation invariant function $f(\cdot)$ can be applied to aggregate and project k instance-level feature vectors into a joint bag-level representation. We use a multilayer perceptron-based attention module as $f(\cdot)$ [ITW18], which produces a combined bag-level feature vector $\mathbf{v}'$ and a set of attention values representing the relative contribution of each instance as defined in Equation (5.2).

$$\begin{aligned} \mathbf{v}' &= f(\mathbf{V}) = \sum_{i=1}^{k} \alpha_i \mathbf{v}_i \\ \alpha &= \mathrm{Softmax}[\mathbf{u}^T \tanh(\mathbf{W}\mathbf{V}^T)] \end{aligned} \tag{5.2}$$

where $\mathbf{V} \in \mathbb{R}^{k \times d}$ contains the feature vectors for k tiles, $\boldsymbol{u} \in \mathbb{R}^{h \times 1}$ and $\boldsymbol{W} \in \mathbb{R}^{h \times d}$ are parameters in the attention module, and h denotes the dimension of the hidden layer. The slide-level prediction can be obtained by applying a fully connected layer to the bag-level representations $\mathbf{v}'$. Both the CNN feature extractor and the attention-based aggregation function are differentiable and can be trained end-to-end using gradient descent. The attention module not only provides a more flexible way to incorporate information from instances, but also enables us to localize informative tiles.

However, this framework encounters similar problems as other saliency detection models [ZWF18, HJW18, SL17]. In particular, as pointed out in [ITW18], instead of detecting the all informative regions, the learned attention map can be highly sparse with very few positive instances having large values. This issue may be caused by the underlying MIL assumption that only one positive instance needs to be detected for a positive bag. Though it might not affect the performance of the cancer detection stage model, this can affect the performance of our classification stage model, which relies on informative tiles selected by the learned attention map. To encourage the model to select more relevant tiles, we use an instance dropout method similar to [SL17, SYS18]. Specifically, instances are randomly dropped during the training, while all instances are used during model evaluation. To ensure the distribution of inputs for each node in the network remains the same during training and testing, pixel values of dropped instances are set to be the mean RGB value of the dataset [SL17, SYS18]. This form of instance dropout can be considered a regularization method that prevents the network from relying on only a few instances for bag-level classification.

5.2.3 Attention-based tile selection

An intuitive approach to localize suspicious regions with learned attention maps is to use the top q percent of tiles with the highest attention weights. However, the percentage of cancerous regions can vary across different cases. Therefore, using a fixed q may cause over selection for slides with small suspicious regions and under selection for those with large suspicious regions. Moreover, this method relies on an attention map, which in this context is learned without explicit supervision at the pixel- or region-level. To address these challenges, we incorporate information embedded in instance-level representations by selecting informative tiles from clusters. Specifically, instance representations obtained from the MIL model are projected to a compact latent embedding space using principle component analysis (PCA). We then perform K-means clustering to group instances with similar semantic features based on their PCA transformed instance-level representations. The relevance of each cluster $\bar{\alpha}_s$ can be determined by the average attention weights of tiles within it as defined by $\bar{\alpha}_s = \frac{1}{m}\sum_{j=1}^{m}\alpha_j$. The intuition is that clusters that contain more relevant information for slide classification should have higher average attention weights. For example, in a cancer-positive slide, clusters consisting of cancerous glands should have higher attention weights compared to those with benign glands and stromal regions. Finally, we can determine the number of tiles to extract from each cluster based on the $\bar{\alpha}_s$ and the total number of tiles.

5.2.4 Multi-resolution WSI classification

Different from most medical imaging modalities, WSIs typically contain billions of pixels, which make them practically impossible to feed into GPU memory directly at full resolution. Though the size of WSIs is enormous, most regions typically do not contain relevant information for slide classification, such as stroma and benign glands. Pathologists tend to analyze the entire slide at a relatively low resolution, usually at 5x, to find suspicious regions and then switch to higher magnification in these areas to render a final diagnosis. Our proposed MRMIL model mimics this process, containing two stages as shown in Figure 5.1. The detection stage model, which consists of an attention-based MIL with instance dropout, is trained with all tiles extracted at a lower magnification (*i.e.*, at 5x) to differentiate benign and malignant slides and generate attention maps. The attention-based clustering method is applied to select relevant tiles for the classification stage model. Selected tiles are extracted at the same location, but at a higher magnification (*i.e.* at 10x) and fed into the MIL network for cancer grade prediction.

5.3 Experiment

5.3.1 Dataset and data preprocessing

5.3.1.1 Dataset

Our dataset contains 20,229 slides from prostate needle biopsies from 830 patients pre- or post-diagnosis (IRB16-001361). Slides' labels extracted from their corresponding pathology reports. There are no additional fine-grained annotations at the pixel- or region-level for this dataset. Additionally, we did not rely on any pre-trained tissue, epithelium, or cancer segmentation networks, and did not perform extensive manual curation to exclude slides with artifacts such as air bubbles, pen markers, dust, *etc.* We randomly divided the dataset into 70% for training, 10% for validation, and 20% for testing, stratifying by patient-level

Table 5.1: Number of slides for each Grade group.

	Train	Validation	Test	Total
No. BN slides	3,225	2,579	5,355	11,159
No. GG 1 slides	3,224	412	807	4,443
No. GG 2 slides	1,966	307	587	2,860
No. GG 3 slides	648	95	148	891
No. GG 4 slides	306	17	129	452
No. GG 5 slides	269	67	88	424
No. Patients	575	86	169	830

Gleason grade group (GG) determined by the highest GG in each patient's set of biopsy cores. This process produced a test set with 7,114 slides from 169 patients and a validation set containing 3,477 slides from 86 patients. From the rest of the dataset, we balanced sampled benign (BN), low grade (LG), and high grade (HG) slides. Table 5.1 shows more details on the breakdown of slides.

5.3.1.2 External dataset

We evaluated our models on a public prostate dataset, SICAPV1, collected by the Hospital Clínico Universitario de Valencia, which contains 512 × 512 tiles at 10x extracted from 79 slides of prostate needle biopsies with 50% overlapping [ELC19]. 19 of these slides are benign, and the rest are malignant.

5.3.1.3 Data preprocessing

The majority of regions on WSIs are background. Thus, we converted slides downsampled at their lowest available magnification compressed in the .svs file into HSV color space and thresholded on the hue channel to produce tissue masks. Morphological operations such as dilation and erosion were used to fill in small gaps, remove isolated points, and further refine tissue masks. Examples on the preprocessing pipeline are presented in the Appendix 5.7.5. We then extracted tiles of size 256×256 at 10x from the grid with 12.5% overlap. Tiles that contain less than 60% tissue were discarded from analysis. The number of tiles per slide ranges from 1 to 1,273, with an average of 275. To account for stain variability, we used a color transfer method [RAG01] to normalize tiles extracted from the slide. The scanning objective was set at 20x (0.5 μm per pixel). We downsampled tiles to 5x for the detection stage model development.

For external dataset, we divided the 512×512 tiles into 4 non-overlapping 256×256 sub-tiles, in order to match the input size of our models. The same stain normalization [RAG01] was applied.

5.3.2 Implementation details

We used VGG11 with batch normalization (VGG11bn) [SZ14] as the backbone for the feature extractor in the MRMIL model for both detection stage and classification stage. A 1×1 convolutional layer was added after the last convolutional layer of VGG11bn to reduce dimensionality and generate $k \times 256 \times 4 \times 4$ instance-level feature maps for k tiles. Feature maps were flattened and fed into a fully connected layer with 256 nodes, followed by ReLU and dropout layers. This produced a $k \times 256$ instance embedding matrix, which was forwarded into the the attention module. The attention part, which generated a $k \times n$ attention matrix for n prediction classes, consisted of two fully connected layers with dropout, tanh nonlinear activations, and a softmax layer. Instance embeddings were multiplied with attention weights, resulting in a $n \times 256$ bag-level representation, which was flattened and input into the final classifier. The probability of instance dropout was set to 0.5 for both model stages. Detailed model architectures were shown in the Appendix Table 5.4 and Table 5.5.

The CNN feature extractor was initialized with weights learned from the ImageNet dataset [DDS09]. After training the attention module and the classifier with the feature extractor frozen for three epochs, we trained the last three VGG blocks together with the attention module and classifier for 97 epochs. The initial learning rates for the feature extractor were set at 1×10^{-5} and 5×10^{-5} for the attention module and the classifier, respectively. The learning rate was decreased by a factor of 10 if the validation loss did not improve for the last 10 epochs. We used the Adam optimizer [KB14] and a batch size of one. Detection stage and classification stage models were trained separately using the same training hyperparameter (*e.g.,* learning rate, batch size and *etc.*.).

For clustering-based region selection, we projected $k \times 256$ instance embedding matrix to $k \times 32$ with PCA, and utilized K-Means clustering to group tiles. The number of clusters was set to be 3 to encourage tiles to be grouped into LG, HG and BN clusters.

Hyper-parameters were tuned on the validation set. We further extended our MRMIL model for GG prediction. The cross entropy loss weighted by reversed class frequency was utilized to address the class imbalance problem. Hyperparameters were selected using the validation set. Models were implemented in PyTorch 0.4.1 [PGC17], and trained on an NVIDIA DGX-1.

5.3.3 Evaluation metrics

As our test dataset contained over 75% benign slides, accuracy (Acc) alone is biased metric for model evaluation. In addition, we used the AUROC and AP computed from ROC and precision and recall (PR) curves, respectively. For cancer grade classification, we measured the Cohen's Kappa (κ), $\kappa = \frac{p_o - p_e}{1 - p_e}$. p_o is the agreement between observers and p_e is the probability of agreement by chance. All metrics were computed using the scikit-learn 0.20.0 package [PVG11].

5.3.4 Model visualization

In addition to quantitative evaluation metrics, interpretability is important in developing explainable machine learning tools, especially for medical applications. In order to have a better understanding of our model predictions, we performed t-Distributed Stochastic Neighbor Embedding (t-SNE) [MH08] of learned bag-level representations for both stage models. t-SNE approach projects high dimensional data into two or three dimensional space where similar data stay closer. t-SNE first produces a joint probability distribution so that similar samples are assigned higher probabilities. It then minimizes the KL-divergence between joint distributions of the projected low dimensional embedding and the original high dimensional data.

To adopt the t-SNE method for visualization of bag-level features, for each slide we utilized the flattened $n \times 256$ feature vector that is aggregated from tile-level representations and is used as the input to the final classification layer. The learning rate of t-SNE was set at 1.5×10^2, and the perplexity was set at 30.

The saliency map produced by the attention module in the MRMIL model only demonstrated the relative importance of each tile. To further localize discriminative regions within tiles, we utilized Gradient-weighted Class Activation Mapping (Grad-CAM) [SCD17]. Concretely, given a trained MRMIL model and a target class c, we first retrieved the top k tiles with the highest attention weights, which were fed to the model. Assume $\boldsymbol{o}_c$ was the model output before the softmax layer for class c, gradients of $\boldsymbol{o}_c$ $w.r.t$ activations $\mathbf{A}^l$ of l-th feature map in the convolutional layer were obtained through backpropagation. Global average pooling over m regions was utilized to generate weights that represent the importance of $w \times h$ feature maps. Weighted combinations of d dimensional feature maps then determined the attention distribution of m regions[SCD17] for predicting the target class c as defined in Equation (5.3).

$$
\begin{aligned}
\boldsymbol{\theta}_l^c &= \frac{1}{Z}\sum_{i\in w}\sum_{j\in h}\frac{\partial \boldsymbol{o}_c}{\partial \mathbf{A}_{i,j}^l} \\
\boldsymbol{\alpha}^c &= \mathrm{ReLU}(\sum_{l=1}^{d}\boldsymbol{\theta}_l^c\mathbf{A}^l)
\end{aligned}
\tag{5.3}
$$

where $Z = w \times h$ is the normalization constant. The ReLU function removes the effect of pixels with negative weights, since they don't have a positive influence in predicting the given class. $\boldsymbol{\alpha}^c$ represents obtained "visual explanation maps" for each image.

5.3.5 Model comparison

5.3.5.1 Handcrafted features

We converted input tiles at 10x into HSV color space and thresholded on the H channel to get tissue masks. Then we utilized the PyRadiomics package [VFP17] to extract 90 features for each tile, including 16 first-order statistics, 23 gray level co-occurrence matrix-based, 16 gray level run length matrix-based, 16 gray level size zone matrix-based, 5 neighbouring gray tone difference matrix-based, and 14 gray level dependence matrix-based features. The maximum pooling was applied to aggregate tile-level features, which were fed into the final slide-level classifier. We experimented with Xgboost [CG16] and random forest (RF) [LW02] classifiers. Grid search with 3-fold cross validation was used to select hyperparameters for classifiers.

5.3.5.2 MIL model by Campanella *et-al.* [CHG19]

We compared our model with the related recent work [CHG19], which also trained slide classification models with only slide-level labels in the MIL framework. Different from our model, they utilized an instance-level MIL approach. The CNN model was trained on top k tiles with high probability after applying the partially trained model, and this process is iterated for certain epochs. Then they utilized the RNN model to aggregate features from top k tiles for final classification. We used the implementation provided by [CHG19] and hyperparameters reported in the paper to re-train the model on our dataset. For fair comparison, we also used the VGG11bn as backbone for feature learning.

5.3.5.3 MIL with different aggregation methods

To evaluate the power of using a trainable attention module, we compared our model with MIL models with pre-defined aggregation functions. Specifically, instead of using the attention module to aggregate tile-level features to slide-level representations, we experimented with two commonly used aggregation methods: maximum pooling and mean pooling aggregation.

5.3.5.4 Single stage model

To evaluate the effectiveness of the multi-resolution model in cancer grade prediction, we compared our model with a model trained with all extracted tiles at 5x only, referred as Single stage.

5.3.5.5 Blue ratio selection

Blue ratio (Br) image conversion, as defined in Equation (5.4), can accentuate the blue channel of a RGB image and thus highlight proliferate nuclei regions [CLP12].

$$\mathrm{Br} = \frac{100 \times B}{1 + R + G} \times \frac{256}{1 + R + G + B} \tag{5.4}$$

where R, G, B are the red, green and blue channels in the original RGB image.

Br conversion is one of the most commonly used approaches to detect nuclei [CLP12, SCR18] and select informative regions from large-scale WSIs [TAO17, AC18, LSF19]. To evaluate the attention-based ROI detection, we replaced the first stage cancer detection model with the Br conversion to select the top $q = 25\%$ tiles with highest average Br values, referred to as *br selection*.

5.3.5.6 Without instance dropout

In this experiment, denoted as *w/o instance dropout*, we investigated whether instance dropout could improve the integrity of learned attention map of the vanilla attention MIL model [LLG19] and lead to better performance.

5.3.5.7 Attention-only selection

Instead of selecting informative clusters, we only utilized the attention map by choosing the top $q = 25\%$ tiles with the highest attention values as the input for the second stage model in the *att selection* experiment.

5.4 Results

Figure 5.2 shows both ROC and PR curves for the detection stage cancer models trained at 5x. The detection stage model in the MRMIL obtained an AUROC of 97.7% and an AP of 96.7% on our internal test set. On the external dataset, it achieved an AUROC of 99.4% and an AP of 99.8%. The model trained without using the instance dropout method yielded a slightly lower AUROC and AP on both internal and external datasets.

Since our dataset does not have fine-grained annotations, we visualized generated attention maps and compared them with pen markers annotated by pathologists during diagnosis. We masked out markers as mentioned in §5.3.1, thus they were not utilized for model training. Figure 5.3 presents the comparison between attention maps learned from models with and without using instance dropout during training.

To further localize suspicious regions within a tile and better interpret model predictions,

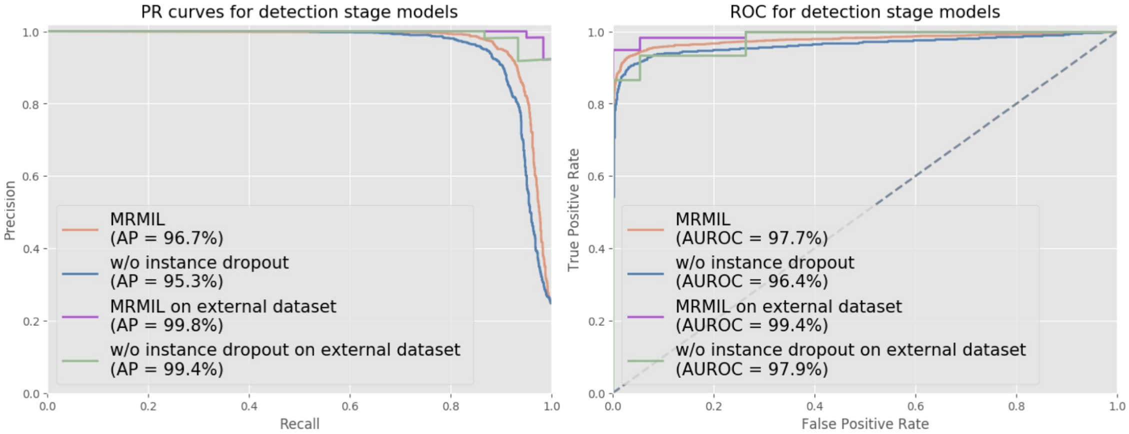

Figure 5.2: ROC and PR curves for detection stage models on our test set and external dataset. In the detection stage, models were trained to distinguish malignant and benign slides with all tiles extracted from slides at 5x.

we applied Grad-CAM on the first detection stage MIL model as shown in Figure 5.4. We generated Grad-CAM maps for not only true positives (TP), but also false positives (FP) to understand which parts of the tile led to false predictions. We selected three tiles with highest attention weights from each slide for visualization.

The MRMIL model projects input tiles to embedding vectors, which are aggregated and form slide-level representations. The t-SNE method enables high dimensional slide-level features to be visualized at a two dimensional space as demonstrated in Figure 5.5. Figure 5.5 (A) is the t-SNE plot for the detection stage model and (B) presents bag-level features produced by the classification stage model with selected high resolution tiles as inputs.

Table 5.2 shows model performances on BN, LG, HG classification. Details for models in each experiment are listed in the Table 5.3. The proposed MRMIL outperformed all baseline models and achieved the highest Acc of 92.7% and κ of 81.8% as shown in row 12. Models with handcrafted features only obtained about 57% κ as demonstrated in row 3 and 4 in the Table 5.2. As shown in row 5, the model by Campanella *et-al.* [CHG19] got 4% lower κ compared with our MRMIL model. Models with simple mean and maximum

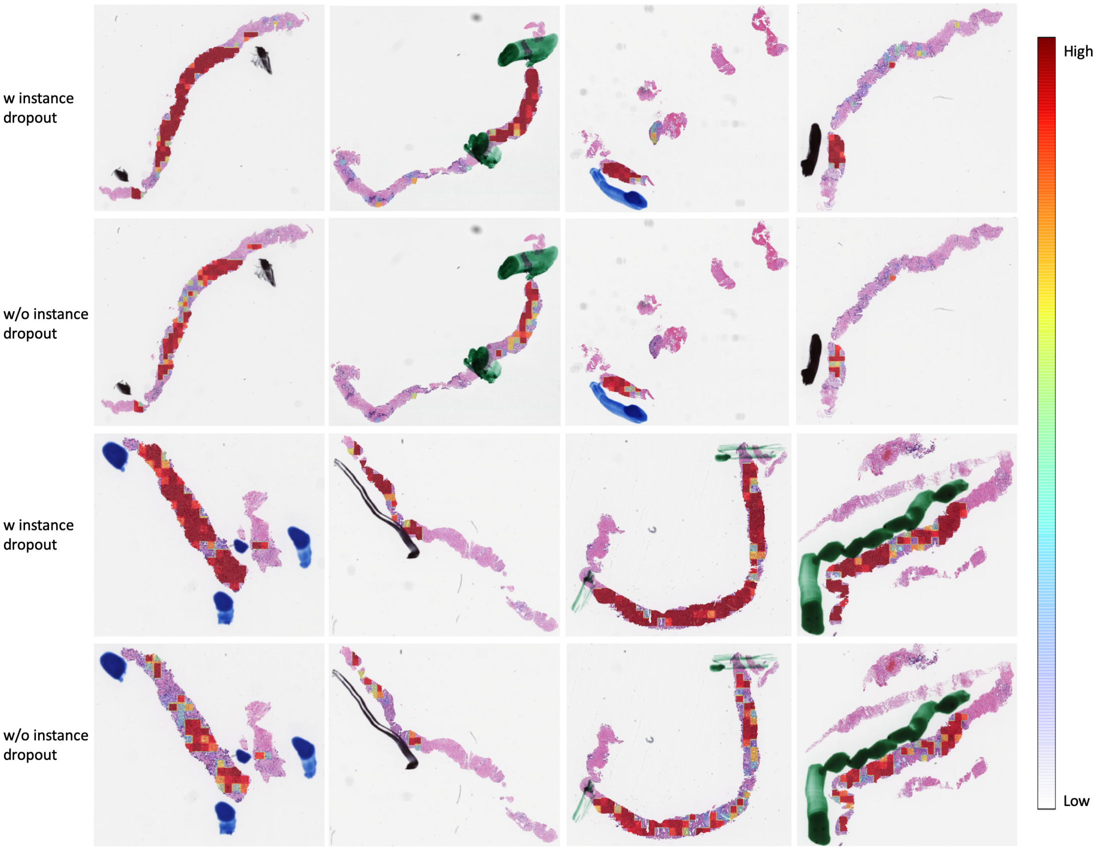

Figure 5.3: WSIs overlaid with attention maps generated from the first stage cancer detection model. Pen markers as mentioned in §5.3.1 indicate cancerous regions. The first row shows attention maps from the model with instance dropout, while the second row is from the model without using instance dropout. Figures are best viewed in color.

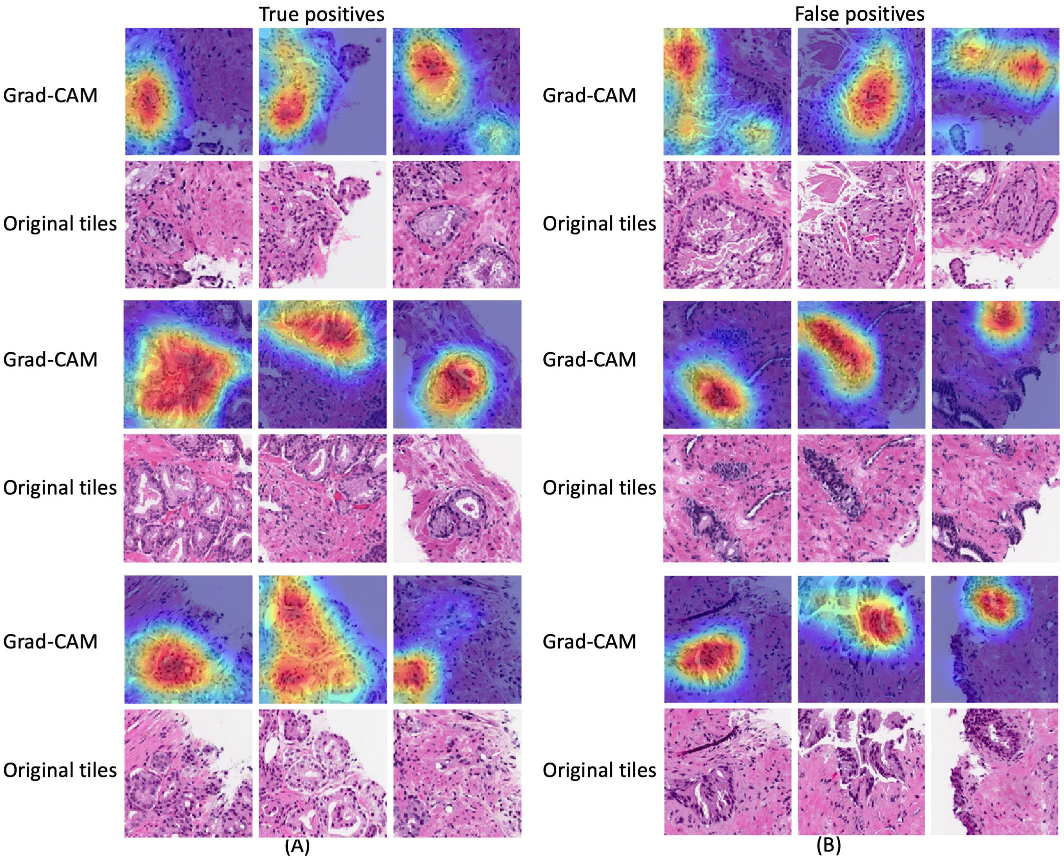

Figure 5.4: Visualization of discriminative regions within tiles for TP and FP predictions. For each slide, we selected the top three tiles with the highest attention weights from the model, which were then forwarded to the model to generate activations and gradients for Grad-CAM.

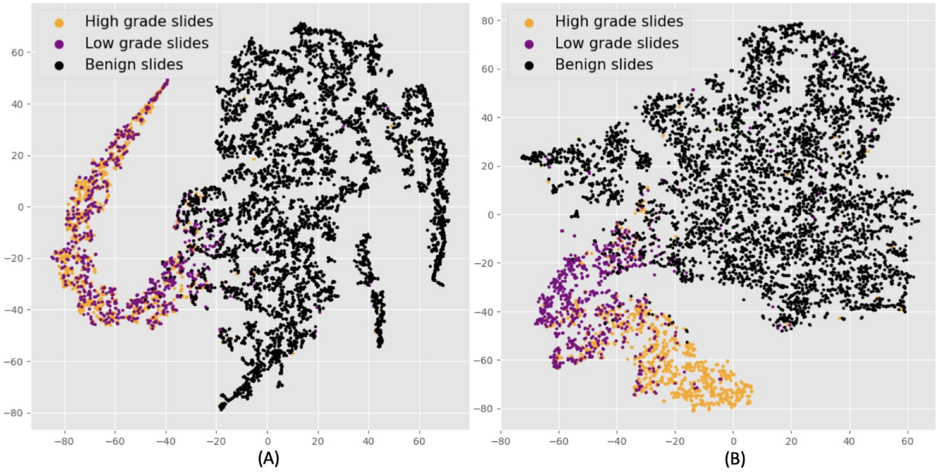

Figure 5.5: t-SNE visualization of slide-level features. Black dots denote benign, purple dots indicate LG, and orange dots represent HG slides. (A) is the slide-level representations from the detection stage model. There is distinct separation between benign and cancerous slides. (B) shows the slide-level features from the classification stage model. We can see a better separation between LG and HG slides.

pooling aggregation methods also achieved lower performance than the MRMIL model as reported in row 6 and 7. Row 8 to 11 demonstrated results on ablation study of the MRMIL model. The single stage attention MIL model trained at 5x achieved 76.3% κ. The br selection that relied on the Br image for tile selection only obtained an Acc of 90.8% and a κ of 76.0%. The w/o instance dropout model, got roughly 4% lower κ and 2% lower Acc compared with the MRMIL model. In addition, we combined LG and HG predictions from the classification model and computed the AUROC and AP for detecting cancerous slides. For instance, by zooming in on suspicious regions identified by the detection stage model, the MRMIL achieved an AUROC of 98.2% and an AP of 97.4%, both of which are higher than the detection stage only model. We present the confusion matrix for the MRMIL model on GG prediction in Figure 5.6. The MRMIL model obtained an accuracy of 87.9%, a quadratic κ of 86.8%, and a κ of 71.1% for GG prediction.

Table 5.2: Model performance on BN, LG, and HG slides classification. Cohen's Kappa and overall accuracy are reported in the table. To evaluate model performances on detecting malignant slides, probabilities for LG and HG are combined. AUROC and AP for cancer detection are also included in the table.

	BN, LG, HG Classification		Cancer Detection	
Experiment Name	Cohen's Kappa (%)	Acc (%)	AUROC (%)	AP (%)
Handcrafted + RF	57.0	81.5	93.1	83.9
Handcrafted + Xgboost	55.9	80.9	93.3	83.9
Campanella *et al* [CHG19]	77.2	90.7	98.3	97.3
Mean aggregation	77.1	90.8	97.9	96.6
Max aggregation	79.5	91.9	97.4	96.3
Single stage	76.3	90.5	97.4	95.8
Br selection	76.0	90.8	95.9	94.3
W/o instance dropout	77.3	91.0	97.3	96.0
Att selection	80.7	92.4	**98.4**	**97.4**
MRMIL	**81.8**	**92.7**	98.2	**97.4**

Table 5.3: Details on models for experiments.

Experiment Name	Model Details
Handcrafted + RF	90 radiomics features + RF at 10x
Handcrafted + Xgboost	90 radiomics features + Xgboost at 10x
Campanella *et al* [CHG19]	MIL + RNN at 10x
Mean aggregation	Mean aggregation at 10x
Max aggregation	Max aggregation at 10x
Single stage	Single resolution MIL at 5x
Br selection	Multi-resolution + Br
W/o instance dropout	Multi-resolution + Att
Att selection	Multi-resolution + Att + instance dropout
MRMIL	Multi-resolution + Att + instance dropout + clusters

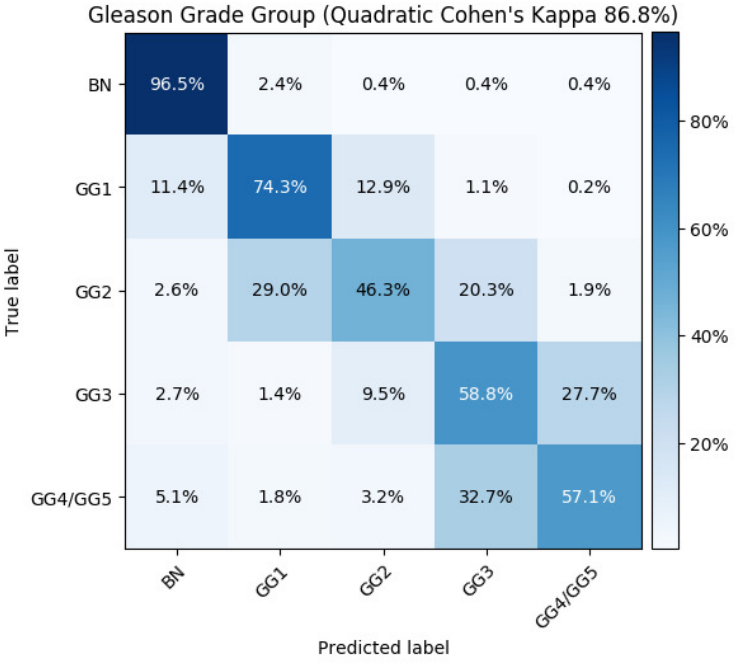

Figure 5.6: Confusion matrix for Gleason grade group prediction.

5.5 Discussion

Our detection stage model achieved promising results on both an internal test set and an external dataset, which demonstrates the generalizability of the model. One potential explanation for slightly better performances on external dataset is that our independent test set is relatively large (*i.e.* 7114 slides from 830 patients.) and is collected from clinical database without any data curation.

Handcrafted features-based models performed relatively well on differentiating benign and malignant slide with an AUC of 93.3%, however, they obtained much lower κ on the hard task of classifying LG, HG and BN slides. The model proposed by Campanella *et al* [CHG19] first used an instance-based MIL approach, which considered tiles with highest probabilities as having the same label as the corresponding slide, and then utilized the RNN model to aggregate representations from top tiles for slide classification. In contrast, our model used a more flexible attention aggregation method that can detect discriminative tiles and combine tile-level features in the same time. The model [CHG19] achieved comparable performance on detecting cancerous slides with 98.3% AUC and 97.3% AP. Yet, it showed inferior results on predicting LG, HG, and BN classes compared with the MRMIL model. Nagpal *et al* developed a two-stage model for Gleason grade prediction of prostate cancer biopsy slides [NFT20]. Their first stage model, which was trained to provide tile-level Gleason pattern classification, was developed using 114 million labeled tiles from over 1,000 slides of prostatectomies and biopsies. The model obtained a κ of 71.7 % on GG1, GG2, GG3 and GG4/5 prediction. Our model, which does not rely on fine-grained annotations and can be trained with only slide-level labels, achieved a comparable performance ($\kappa = 71.1\%$).

The quality of attention maps from the detection stage model is essential for selecting discriminative regions for the classification stage model. As shown in Figure 5.3, attention maps learned with only weak (*i.e.* slide-level) labels are consistent with cancerous regions identified by pathologists during diagnosis. This demonstrates that our detection stage model not only achieves strong performance in classifying malignant versus benign slides, but also identifies suspicious regions for classification stage models. In addition, the generated attention maps can be integrated into a WSI viewer to potentially help pathologists more quickly localize relevant areas and reduce diagnostic time. Figure 5.3 also shows that the original attention-based MIL model [ITW18] (*i.e.* w/o instance dropout) only focuses on a few most discriminative tiles instead of entire suspicious regions. As reflected in Table 5.2, the w/o instance dropout model obtained a κ of 77.3%, which is about 4% lower than the one trained with instance dropout. Moreover, the performance of the model that relied on the Br image is inferior to the models that utilized attention maps. This demonstrates that areas with the most blue color may not be diagnostic relevant regions and that our attention module is able to extract high-level predictive representations rather than purely color features.

Grad-CAM visualization facilitates understanding of predictions from "black-box" deep learning models, as shown in Figure 5.4. For TP predictions in Figure 5.4 (A), our model captured the most relevant parts in the tile, though some cancerous regions were missed. For example, the first tile on the fourth row contains densely clustered cancerous glands, but the Grad-CAM in the third row only highlighted the most central area, and cancerous glands closer to the boundary were not detected. FP predictions are usually also hard cases for pathologists, with features that resemble prostate cancer. For example, regions highlighted by Grad-CAM in third and fourth rows in Figure 5.4 (B) contain benign glands with increased number of basal cells due to tangential tissue sectioning. Last two rows in Figure 5.4 (B) show the seminal vesicle/ejaculatory duct tissue that form small outpouching glands with amphophilic cytoplasm, which mimic malignant glands. Our model was only trained to detect and grade acinar adenocarcinoma for prostate biopsies. Interestingly, as shown in first two rows in Figure 5.4 (B), the model was able to identify intraductal carcinoma of the prostate gland (IDC-P), which is usually associated with high-stage invasive cancer and adverse prognosis.

From Figure 5.5 (A), we can see that benign slide representations are clustered together on the right and malignant slides form a small cluster on the left. There is no distinct separation between features from LG and HG slides, since the objective of the detection stage model is to classify cancerous versus benign slides. Figure 5.5 (B) shows that features of LG and HG slides generated from the classification stage model form their own distinct clusters, and representations from LG slides lie closer to benign slides in the embedding space.

To quantitatively evaluate our model performance, we performed experiments to understand the contribution of different model components, as summarized in Table 5.2. Using attention maps to select higher resolution tiles improved the κ of the one with br selection by 1%. Instance dropout further boosted the κ by over 3%. The final model MRMIL with all components achieved the highest κ for BN, LG, and HG classification, 98.2% AUROC for detecting malignant slides, and a quadratic κ of 86.8% for GG prediction, which is comparable to state-of-the-art models that require pre-trained segmentation networks [BPB19].

5.6 Summary

In this chapter, we present a novel MRMIL model that consists of a detection stage and a grade classification stage. The model can be trained with weak supervision from slide-level labels and localize cancerous regions. We provided visualization of saliency maps at both the slide- and tile-level, and learned representations to enhance model interpretability. The model was developed and evaluated on a dataset with over 20k prostate slides from 830 patients and an external dataset [ELC19], and achieved promising performance. We believe that these types of models could have multiple applications in the clinic, including allowing pathologists to increase their efficiency, empowering more general pathologists to perform at the level of experts, and performing“second reads” of biopsy slides for quality assurance.

5.7 Appendix

5.7.1 Detailed model architecture

Table 5.4 shows the detailed architecture for the first stage cancer detection stage model, and Table 5.5 shows the architecture for the classification stage model. Two stage models were trained separately.

5.7.2 Blue ratio conversion

Figure 5.7 demonstrates blue ratio conversion method for detecting regions from whole slide images for fine-grained Gleason grade classification. We can see that the method can identify areas with most nuclei, however these region may not be cancerous area (*e.g.,* it may identify regions with large benign glands).

5.7.3 K means clustering

We used PCA to project $n \times 256$ instance-level embedding vectors of n tiles to $n \times 32$ (*i.e.,* the number of components is set to be 32). For K-means clustering, the k was set to be 3 to encourage tiles to be grouped into benign, low-grade and high-grade clusters. Our attention clustering-based selection method was robust to different initializations. Specifically, we reran the K-means clustering with different random seeds for 10 times, and computed mean intersection over union (IoU) for selected tiles. Our method achieved a mean IoU of 97.65%.

Table 5.4: Cancer detection stage model architecture.

Module	Layers	# filters	Filter size	Output size
Input		-	-	3 × 128 × 128
VGG11bn	Conv + BN + ReLU	64	3 × 3	64 × 128 × 128
	Max Pool	64	2 × 2	64 × 64 × 64
	Conv + BN + ReLU	128	3 × 3	128 × 64 × 64
	Max Pool	128	2 × 2	128 × 32 × 32
	Conv + BN + ReLU	256	3 × 3	256 × 32 × 32
	Conv + BN + ReLU	256	3 × 3	256 × 32 × 32
	Max Pool	256	2 × 2	256 × 16 × 16
	Conv + BN + ReLU	512	3 × 3	512 × 16 × 16
	Conv + BN + ReLU	512	3 × 3	512 × 16 × 16
	Max Pool	512	2 × 2	512 × 8 × 8
	Conv + BN + ReLU	512	3 × 3	512 × 8 × 8
	Conv + BN + ReLU	512	3 × 3	512 × 8 × 8
	Max Pool	512	2 × 2	512 × 4 × 4
Instance embedding	Conv	256	1 × 1	256 × 4 × 4
	FC + ReLU + Dropout	-	-	256
Attention module	FC + Tanh + Dropout	-	-	512
	FC	-	-	1
Classifier	FC	-	-	1

Table 5.5: Cancer classification stage model architecture.

Module	Layers	# filter	Filter size	Output size
Input		-	-	3 × 256 × 256
VGG11bn	Conv + BN + ReLU	64	3 × 3	64 × 256 × 256
	Max Pool	64	2 × 2	64 × 128 × 128
	Conv + BN + ReLU	128	3 × 3	128 × 128 × 128
	Max Pool	128	2 × 2	128 × 64 × 64
	Conv + BN + ReLU	256	3 × 3	256 × 64 × 64
	Conv + BN + ReLU	256	3 × 3	256 × 64 × 64
	Max Pool	256	2 × 2	256 × 32 × 32
	Conv + BN + ReLU	512	3 × 3	512 × 32 × 32
	Conv + BN + ReLU	512	3 × 3	512 × 32 × 32
	Max Pool	512	2 × 2	512 × 16 × 16
	Conv + BN + ReLU	512	3 × 3	512 × 16 × 16
	Conv + BN + ReLU	512	3 × 3	512 × 16 × 16
	Max Pool	512	2 × 2	512 × 8 × 8
Instance embedding	Conv	256	1 × 1	256 × 8 × 8
	FC + ReLU + Dropout	-	-	256
Attention module	FC + Tanh + Dropout	-	-	512
	FC	-	-	3
Classifier	FC	-	-	3

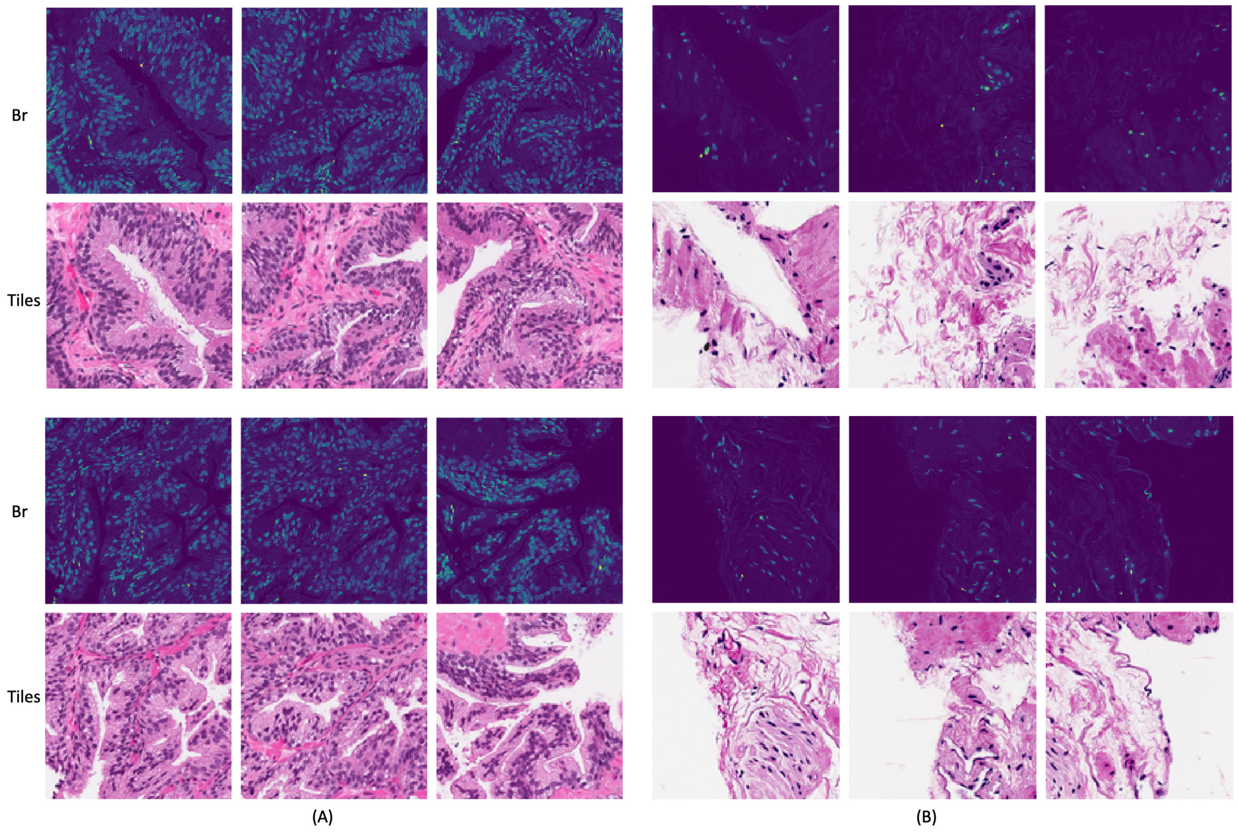

Figure 5.7: Br conversion. We performed Br conversion for slides at 5x. The first two rows demonstrate tiles from a benign slide and the bottom two show ones from a malignant slide. (A) are 3 tiles with the highest average tile-level Br values, and (B) are ones with the lowest Br values. We can see that the br conversion is able to highlight regions with most nuclei.

Table 5.6: Model performance on BN, LG, and HG slides classification

MRMIL with different backbones	BN, LG, HG classification		Cancer detection	
	Cohen's Kappa (%)	Acc (%)	AUROC (%)	AP (%)
VGG11bn	81.8	92.7	98.2	97.4
VGG13bn	79.9	92.0	97.8	96.9
ResNet34	78.7	91.6	96.9	95.3

5.7.4 Different CNN architectures for the feature extractor

We performed experiments to evaluate our MRMIL model performances with different network backbones. Experiments were performed by replacing the feature extractor in both model stages with different CNN architectures. As shown in the Table 5.6, the VGG11bn achieved the best performance. Our model performances were affected around 2% by using different backbones.

5.7.5 Tissue region detection

Despite the enormous size of WSIs, tissue may only occupy small regions with most areas being background. We developed a tissue detection pipeline to remove background and pen markers from analysis. The slide is first converted into the hue saturation and value (HSV) color space where the H channel controls the color, the V channel determines the brightness of the color and the S channel specifies the saturation of color. We threshold on the H channel to remove pen markers and the threshold on the S channel to remove white background. Morphological operations are then applied on the generated binary mask to remove small objects and holes. Figure 5.8 shows examples from the tissue detection pipeline for slides from prostate biopsy. Additional examples for tissue detection on slides from radical prostatectomy are presented in Figure 5.9.

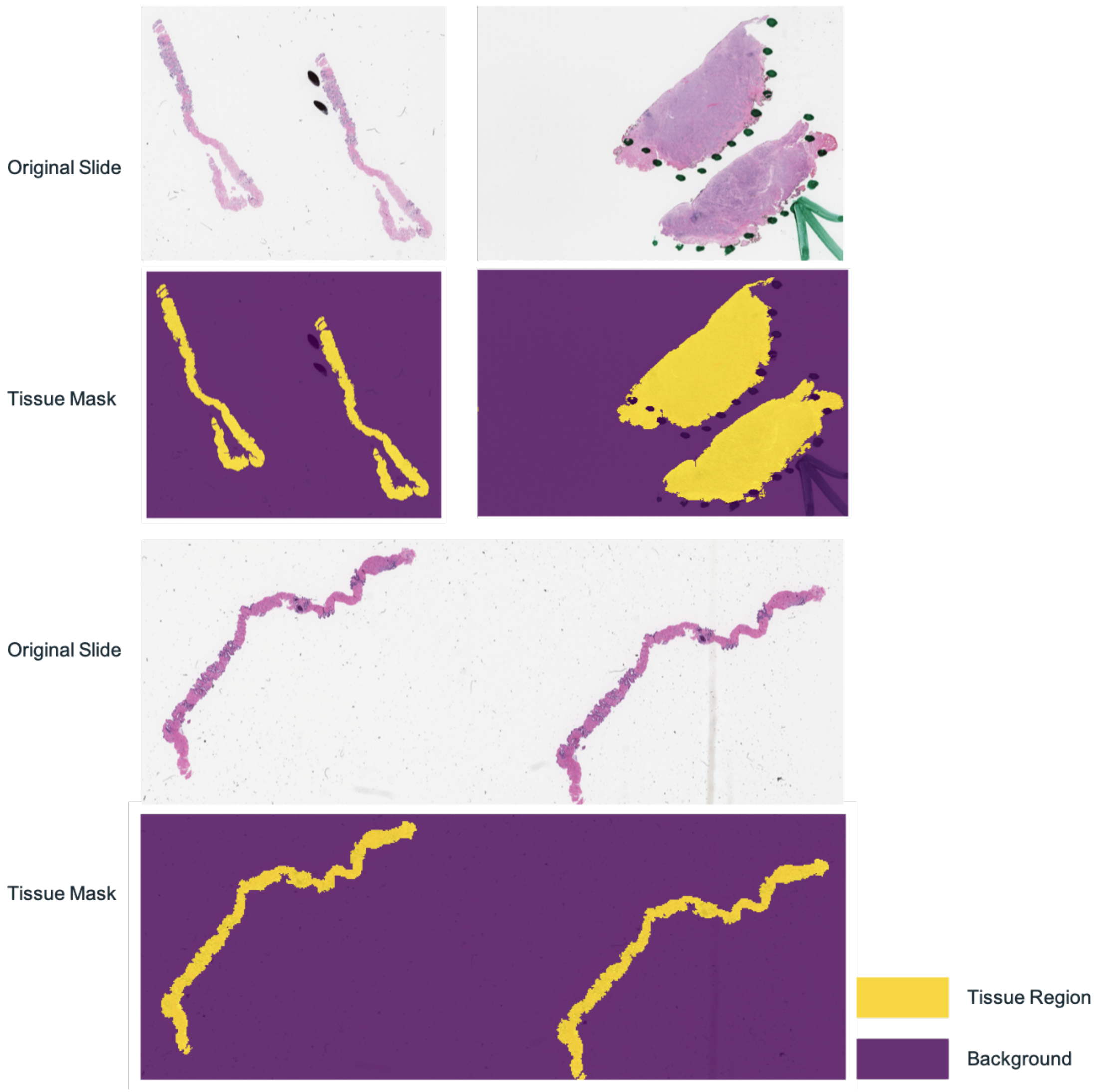

Figure 5.8: Examples of tissue detection results for biopsy slides. The processing pipeline is able to separate background and pen markers from tissue regions. This could help downstream models focus on the relevant tissue regions.

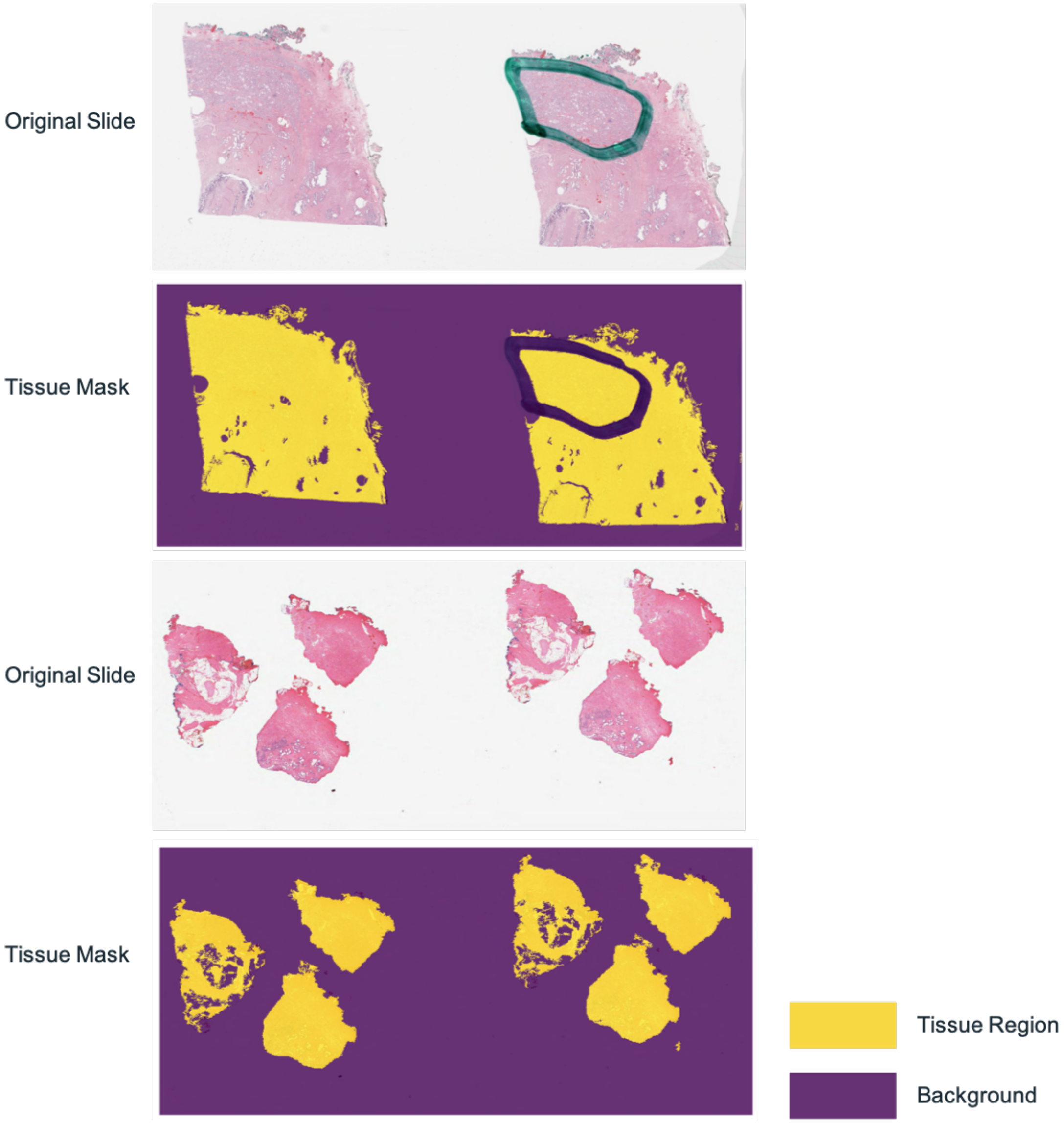

Figure 5.9: Additional examples of tissue detection results for prostatectomy slides.

CHAPTER 6

Progression-free Analysis with Pathomic Features

Chapter 3 and 4 present computer aided diagnosis (CAD) tools that can produce pixel-wise Gleason pattern predictions for a given tile from whole slide images (WSIs). Instead of focusing on tile-level analysis, Chapter 5 details a CAD system, which can detect diagnostically relevant areas and produce slide-level Gleason grade group (GG) predictions. Though currently Gleason grading system plays an essential role in prostate cancer diagnosis and treatment planning, heterogeneous outcomes may be observed in patients even within the same Gleason score [MWH16, MIW19]. For example, the study performed by McKenney *et al* suggested more histologic patterns should be considered and regrouped in the current Gleason grading system (*e.g.,* cases with cribriform G4 patterns have a higher risk comparing with those with G4 with poorly formed glands).

Rich visual and sub-visual patterns embedded in WSIs could provide important information for disease prognosis. Due to the scarcity of data with progression labels, most previous works on prostate progression analysis utilized handcrafted features or CNN models pre-trained on natural image datasets [DDS09]. Moreover, relying on manually selected tiles or pre-defined pooling functions (*e.g.,* maximum and mean pooling) to convert tile-level information to case-level predictions were mainly used in previous studies. Challenges remain on how to effectively extract discriminative representations and combine tile-level feature vectors for slide-level representations. In this Chapter, we further investigate the capability of quantitative self-supervised learning features, attention and spatial-aware aggregation in a computer-aided progression system for progression-free survival (PFS) prediction for prostate patients after radical prostatectomy.

Details on different components in our progression prediction system are presented in §6.1. In §6.2 the dataset used in this work is reviewed. Experiments and model implementation are illustrated in §6.3. Results for experiments are discussed in §6.4. Finally, in §6.5 and §6.6 we summarize advantages and limitations of our model, and discuss potential directions of future work.

6.1 Methods

Our progression prediction system contains two main parts: 1) a tumor region detection model as detailed in §6.1.1, which facilitates the selection of informative tiles; 2) a deep survival model, which aggregates discriminative tile-level features to predict the case-level survival.

We experiment with two different approaches for tile-level representation learning: a) handcrafted texture features; b) momentum contrast self-supervised learning model (MoCo) [HFW20]. To summarize tile-level features, we develop two aggregation methods: 1) an attention MIL-based method [LLG19]; 2) a graph convolutional neural network-based (GCN) method. Figure 6.1 shows an overview of our system. Details of each module are introduced in the following sections.

6.1.1 Tumor region detection

One challenge for analyzing large-scale WSIs is to identify suspicious regions to sample tiles. Previous approaches relied on manually circled cancerous regions, which are expensive to collect [VCC17, DAH18, YZB16, CLL20, LJE19, LRW18, NFT20, NFL19].

In this work, we utilize a pre-trained tile-level CNN to produce cancer probability map for WSIs. k tiles with highest cancer probability are selected for survival analysis. Training

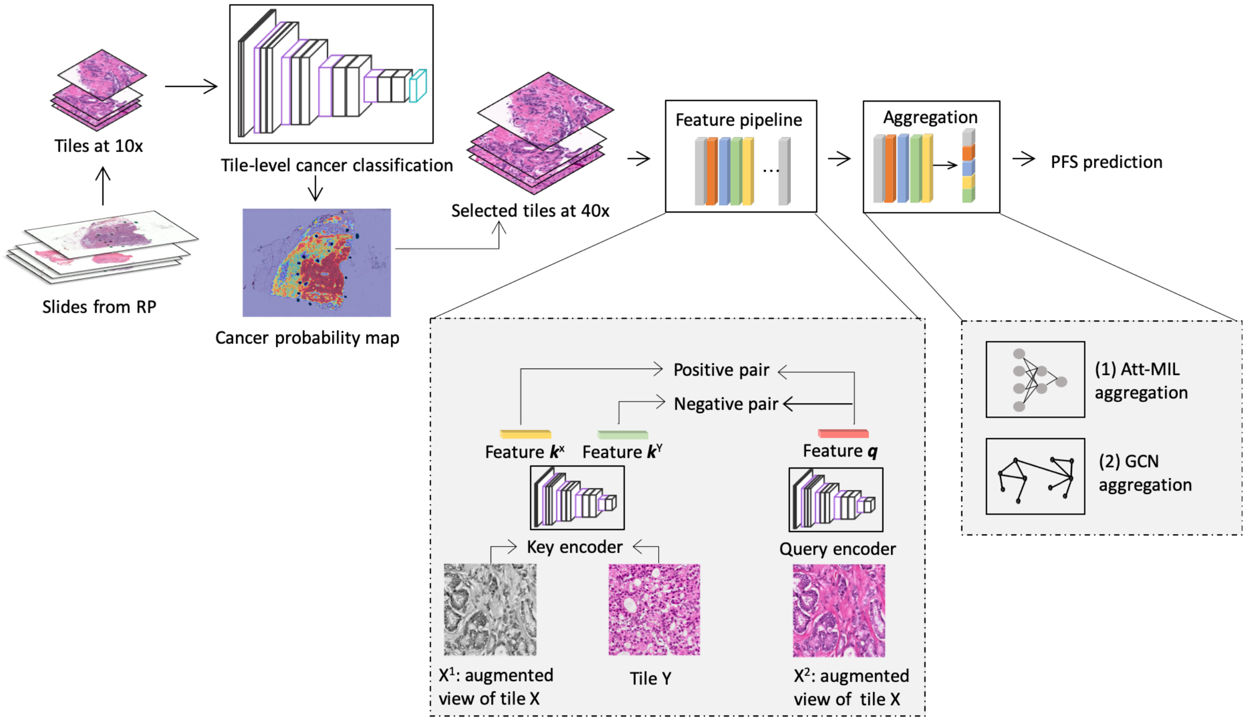

Figure 6.1: Overview of the proposed PFS prediction system. The system mainly consists of two stages: 1) tumor detection stage, where a tile-level cancer classification model is applied to produce cancer probability map for each slide; 2) PFS prediction stage, where representations are extracted from selected tiles and aggregated into a slide-level feature vector for PFS prediction. The figure shows an example for self-supervised learning-based feature extraction with the MoCo model. Rather than using task specific labels, MoCo utilizes contrastive loss, which encourages smaller distances between features from augmented views of the same tile, but larger distance among other tiles.

a traditional tile-level CNN may require a large amount of expensive samples labeled at local-level (*e.g.,* region-level or pixel-wise). To deal with the label scarcity challenge, we exploit a large-scale dataset with weakly-labeled biopsy slides and a small dataset containing prostatectomy slides with region-level annotations.

Specifically, we first consider the problem of training a model to differentiate benign and malignant tiles with global slide-level labels as a multiple instance learning (MIL) problem. Each slide is modeled as a bag and tiles extracted from the slide are considered as instances. Benign slides should contain only benign tiles. The slide is malignant iff it has a cancerous tile as defined in the Equation 5.1. Thus, if tiles are sorted in descending order according to their probabilities of being cancerous. Top Tiles within a malignant slide will have probabilities closer to 1, while tiles from a benign slide should have probabilities closer to 0. Since tile-level predictions are required, we optimize the tile-level CNN model under an instance-level MIL framework in an iterative way [CHG19].

In step t, the tile-level model from step $t-1$ is applied to produce cancer probability for each tile. Top s tiles from each slide are selected and inherit the corresponding slide-level label. s is set to be 1. Then the tile-level model is further optimized with pseudo tile-level labels. The inference of tile labels and training of the tile-level model are iterated until convergence.

In addition to the large-scale biopsy dataset, we leverage the small prostatectomy dataset with regions of interest circled and graded by pathologists. The tile-level model is then fine-tuned on this dataset using tiles extracted from annotated regions.

6.1.2 Histopathological feature extraction

6.1.2.1 Deep representation learning with MoCo model

Fully supervised CNNs are known to be "data hungry" and require lots of expensive labeled samples. Despite the large size of WSIs, available labels for histopathological datasets are usually sparse and at global-level. For example, a patient may have multiple WSIs, each of which could contain billions of pixels, but only one overall survival label. Self-supervised learning models are proposed to leverage information embedded in unlabeled data. Instead of computing losses between model predictions and task specific labels, contrastive loss has been exploited in some self-supervised learning models to learn task agnostic representations, which can capture the underlying structure of the data [HFW20, CKN20, BHB19, OLV18]. In this work, we adopted a modified momentum contrastive learning model (MoCo) [HFW20] to learn representations embedded in the histopathological images, which may be informative for disease prognosis.

The main assumption for the MoCo model is that features from augmented versions of the same image should be more similar than features from different images. Specifically, images can be projected into feature vectors $k_i, i = 1, 2, 3, ...$ by a key encoding function $f_k(\cdot)$, which can be considered as keys in a dataset. Given the encoded representation $\boldsymbol{q_i}$ of a query image generated by the query encoding function $f_q(\cdot)$, there exists one matched key, which has the largest similarity value to $\boldsymbol{q_i}$. The similarity between a pair of feature vectors can be measured with cosine distance: $\text{sim}(\boldsymbol{q_i}, \boldsymbol{k_i}) = \frac{\boldsymbol{q_i k_i}}{\|\boldsymbol{q_i}\|\|\boldsymbol{k_i}\|}$. The contrastive loss, also referred as the infoNCE [OLV18, HFW20], can be defined by the Equation 6.1.

$$\mathcal{L}_q = -\text{log}\frac{\exp(\text{sim}(\boldsymbol{q_i k_i})/\gamma)}{\exp(\text{sim}(\boldsymbol{q_i k_i})/\gamma) + \sum_{j\neq i}^{N} \exp(\text{sim}(\boldsymbol{q_i k_j})/\gamma)} \tag{6.1}$$

where γ is the temperature parameter. $\boldsymbol{q_i}$ and $\boldsymbol{k_i}$ are feature vectors from different views of the same image i. The numerator denotes the similarity between a positive pair and the denominator represents the sum over similarities of one positive pair and other negative pairs. A positive pair consists of the key and query feature vectors originated from the same image. For example, query and key images can be obtained by applying stochastic data augmentation methods, such as random rotation, cropping, blurring, resizing and *etc*, on the original image. Identifying the matched image is used as the surrogate prediction task in the contrastive loss based self-supervised learning model.

Contrastive learning-based models benefit from a rich set of negative samples. However, this may require a larger batch size, hence, a larger GPU memory. To address this challenge, the MoCo model maintains a running queue of the dictionary containing embeddings of key images [HFW20]. The dictionary is updated continuously with new batches of samples. This enables the number of negative samples to be decoupled from the batch size [HFW20]. To make parameter update tractable for the key encoder f_k, the MoCo model exploits the momentum update, in which the f_q is updated by back-propagation and the f_k is updated gradually by $f_q : f_k = mf_k + (1 - m)f_q$ [HFW20].

Moreover, extensive data augmentation prevents the model from shortcut solutions (*e.g.,* using low-level noisy features such as color to distinguish images) and encourages it to learn high-level, semantic representations. Color distortions is one of the most effective transformations for improving learned representations [CKN20]. In this work, we leverage the property of H&E stained WSIs that the image can be decomposed into two stain channels representing nuclei and stroma areas. Random conversion to H stain channel is included as additional color augmentation method during model training.

High dimensional tiles extracted from WSIs can then be encoded into lower dimensional representations with the trained MoCo model for the down-stream survival prediction.

6.1.2.2 Texture-based features

Texture descriptors measure spatial distribution of intensities. Many previous work in classifying pre-selected patches from prostate WSIs utilized texture-based features. Mosquera-Lopez *et al* provided a review on texture-based CAD tools of prostate cancer [MAV14]. For comparison, we build a pipeline with the Pyradiomics package [VFP17] to extract first-order statistics based and texture-based features. Details about features included in the pipeline are listed in Table 6.1.

Table 6.1: Handcrafted features used for survival analysis. Tiles are first converted into HSV color space. H channel and the binary mask for tissue regions are fed into the Pyradiomics package is used to compute features.

Feature Category	Number of Features
First order statistics	16
Gray-level co-occurrence matrix (GLCM)	23
Gray-level size zone matrix (GLSZM)	16
Gray-level run length matrix (GLRLM)	16
Gray-level dependency matrix (GLDM)	14
Neighbouring gray tone difference matrix	5
Total features	90

6.1.3 Aggregate tile-level features for progression free survival analysis

Each suspicious tile identified by the tumor detection model can be represented by a d dimensional feature vector $\boldsymbol{z}, \boldsymbol{z} \in \mathbb{R}^{d\times 1}$ using the MoCo model or human-engineered feature pipelines as described in §6.1.2. However, the outcome label is at case-level. How to combine tile-level features for case-level predictions is challenging. In this work, we mainly investigate two types of aggregation methods: the attention MIL-based approach (att-MIL) detailed in §6.1.3.1 and the GCN-based method introduced in §6.1.3.2. Assume an aggregated representation $\boldsymbol{z}^{\text{agg}}$ can be generated for each case, negative logarithm of Cox partial likelihood as proposed in the DeepSurv model [KSC16] is utilized as the loss function for model optimization. Given the event e_i is observed for patient i at T_i, the Cox loss can be computed by the Equation 6.2.

$$l(\beta) = -\sum_{i:e_i=1} (h_\theta(\boldsymbol{z}^{\text{agg}}) - log \sum_{j\in R_{(T_i)}} \exp(h_\theta(\boldsymbol{z}^{\text{agg}}))) \tag{6.2}$$

where $j \in R_{(T_i)}$ represents a set of patients with events that haven't occurred at T_i. $h(\cdot)_\theta$ is a non-linear function (*e.g.,* a multi-layer perceptron).

6.1.3.1 Aggregation with attention-based multiple instance learning method

The problem of weakly-supervised survival prediction can be also modeled with a MIL framework, where labels are available for a bag of tiles (*i.e.,* instances), while labels for each tile is unknown. Bag-level MIL method is utilized, since global-level predictions are more important for the case-level survival prediction. Rather than using pre-defined pooling functions, we exploit a trainable attention module to summarize tile-level features into a slide-level representation. The attention module enables weighted aggregation of tile-level features and makes the network to focus on relevant regions.

Specifically, an attention module $g(\cdot)$ is added after the fully connected layer, which produces tile-level representations, to learn weight distribution $\boldsymbol{\alpha} = \alpha_1, \alpha_2, ..., \alpha_N$ for N tiles. The $g(\cdot)$ can be modeled by a multilayer perceptron (MLP). The attention for the k th instance can be defined by the Equation 6.3:

$$\alpha_i = \text{Softmax}[\boldsymbol{U}^T(\tanh(\boldsymbol{W}\boldsymbol{z}_k^T))] \tag{6.3}$$

where $\boldsymbol{U} \in \mathbb{R}^{h \times 1}$ and $\boldsymbol{W} \in \mathbb{R}^{h \times d}$ are learnable parameters, and h is the dimension of the hidden layer. Then each tile can have a corresponding attention value learned from the module. The slide-level embedding $\boldsymbol{z}^{\,agg}$ can be obtained by multiplying learned attentions with instance features. Promising results have been observed with the attention-based MIL method for cancer detection and Gleason grade prediction tasks as shown in Chapter 5. In this chapter, we further extend it for PFS prediction.

6.1.3.2 Aggregation with graph convolution neural network

The attention-based MIL method utilizes trainable attention module, which may fail to model the spatial structure (*i.e.,* topology) of tile-level features. If we consider a slide as a graph and denote tiles as nodes associated with tile-level features, slide-level representations can be also produced using the GCN. The GCN updates features of each node by aggregating information from neighborhood nodes in each graph convolutional layer and creates hierarchical representations by stacking multiple layers. Instead of using a permutation invariant function, GCN takes account of spatial relationships of tiles.

Specifically, the GCN extends the idea of the graph neural network (GNN) [GMS05]. A graph can be denoted by $\mathcal{G}(\mathcal{V}, \mathcal{E})$ consisting of a vertex set $\mathcal{V} = \{v_i\}_{i=1}^{N_\mathcal{V}}$ and edge set $\mathcal{E} = \{e_j\}_{j=1}^{N_\mathcal{E}}$ with feature vector $\boldsymbol{z}_{v_i} \in \mathbb{R}^{d\times 1}$ associated with vertex v_i. Assume $\mathcal{N}_{v_i}^{\mathcal{E}}$ is the neighborhood of the node v_i and mean pooling is used to combine features from neighborhood nodes, a layer of the GCN can be defined by the Equation 6.4:

$$g_\mathcal{E}(v_i) = \text{ReLU}(\boldsymbol{W} \cdot \text{Mean}(\boldsymbol{z}_{v_j} | v_j \in \mathcal{N}_{v_i}^{\mathcal{E}})) \text{ ,where } \boldsymbol{W} \in \mathbb{R}^{d\times n} \text{ is a learnable matrix} \tag{6.4}$$

We use a graph attention network (GAT) [VCC17] as the backbone of our model. The GAT extends and improves the GCN by learning attention scores that indicate the importance of a node's neighborhood features. The attention mechanism is a single-layer feedforward neural network, parametrized by a weight vector $\boldsymbol{\alpha}$. The weighting coefficients computed by the attention mechanism can be expressed by the Equation 6.5:

$$\alpha_{v_i,v_j} = \frac{\exp(\rho(\boldsymbol{a}^T[\boldsymbol{W}\boldsymbol{z}_{v_i} \| \boldsymbol{W}\boldsymbol{z}_{v_j}]))}{\sum_{v_k \in \mathcal{N}_{v_i}^{\mathcal{E}}} \exp(\rho(\boldsymbol{a}^T[\boldsymbol{W}\boldsymbol{z}_{v_i} \| \boldsymbol{W}\boldsymbol{z}_{v_k}]))} \tag{6.5}$$

where $\boldsymbol{a} \in \mathbb{R}^{2d}$ is a trainable weight vector, $\|$ represents concatenation and ρ denotes the LeakyReLU non-linear activation. In the final GCN layer, all remaining node features are combined using a global attention module and form a slide-level representation $\boldsymbol{z}^{\text{agg}}$ that is utilized for PFS estimation.

6.2 Dataset

Three datasets are utilized in this work:

1. UCLA prostate biopsy dataset. The dataset, referred as *dataset-biopsy*, contains 20,229 slides from prostate needle biopsies from 830 patients pre- or post-diagnosis. Slide-level Gleason grade labels are retrieved from pathology reports. The dataset is randomly divided into 70% for training, 10% for validation and 20% for testing, stratified by patient-level Gleason grade group (GG) determined by the highest GG in each patient's set of biopsy cores. More details on label distributions of this dataset are introduced in the §5.3.1. This dataset is mainly utilized to pre-train a tile-level cancer prediction model under the instance-level MIL framework. Slides in dataset-biopsy are scanned with a mixture of scanning objectives: 40x (*i.e.,* 0.25 μm per pixel) and 20x (*i.e.,* 0.5 μm per pixel).

2. Cedars-Sinai prostatectomy dataset. This dataset, referred as *dataset-finetune*, consists of 30 WSIs from prostatectomies of 30 patients [IML18]. These slides were annotated by a pathologist who circled and graded the major foci of tumor as either low-grade (Gleason pattern 3), high-grade (Gleason pattern 4 and 5), or benign (BN) areas. The tile-level cancer prediction model, which is trained with a large weakly annotated biopsy dataset (*i.e.,* only slide-level labels are available), is then fine-tuned on the dataset-finetune that contains region-level annotations. The scanning objective of slides in this dataset is set at 20x. 5-fold cross validation is used for model training and validation.

3. The Cancer Genome Atlas Program prostate dataset (TCGA-PRAD). Models for PFS prediction are developed on the publicly available TCGA-PRAD dataset [CDS18,

AAA15, LLH18]. Clinical data and formalin-fixed paraffin-embedded, H&E stained diagnostic slides are retrieved using the genomic data commons (GDC) data portal. Additional clinical follow-up data are obtained from the standardized dataset provided by the TCGA pan-cancer clinical data resource (TCGA-CDR) [LLH18]. As recommended in the TCGA-CDR, the progression-free interval (PFI) are used as the clinical endpoint [LLH18]. The progression includes biochemical recurrence, locoregional recurrence, distant metastasis and new primary tumor. 399 cases with available diagnostic slides are included in this study. Slides are scanned at 40x in this dataset. 5-fold cross validation stratified by events is utilized for model training and validation. Within each fold, we further split the training data into 80% for training and 20% for validation. Model hyper-parameters are selected using the 20% validation set within the training part of each fold, and model performances are evaluated on left out test data for each fold to avoid overfitting.

6.2.1 Data preprocessing

The same preprocessing pipeline as described in §5.3.1 is exploited for all three datasets to generate tissue masks and facilitate tile extraction. Some examples of generated tissue masks are presented in appendix §5.7.5.

Tiles of size 512×512 at 20x with at least 80% of tissue areas are extracted from the grid for the dataset-biopsy, and sampled from annotated regions for the dataset-finetune. Tiles of size 1024×1024 at 40x with at least 80% tissue areas, which corresponds to 512×512 regions on slides scanned at 20x, are extracted from the grid for the TCGA-PRAD dataset. Tiles from dataset-biopsy and dataset-finetune are downsampled to 10x (256×256) and utilized to train and finetune the tumor detection stage model. The model is applied on tiles from the TCGA-PRAD dataset, which are also downsampled to 10x, to select top 200 tiles with highest cancer probabilities for each slide. The MoCo model is then trained with 512×512 tiles at 40x that are randomly cropped from tiles selected by the tumor detection stage model. During testing, representations are produced from the center cropped tiles.

Different staining protocols, tissue preparation procedures, scanning conditions and *etc* may lead to stain variations. Thus, we utilize the Reinhard [RAG01] color normalization method to normalize generated tiles from all three datasets.

6.3 Experiment

6.3.1 Model training

We used ResNet50 as the backbone for the MoCo model. Data augmentations including random cropping, stain separation, Gaussian blurring and flipping were utilized for the MoCo model training. The feature dimension was set at 128. The number of negative samples in the key dataset was set at 65536. A learning rate of 0.015 and a batch size of 128 were utilized. The SGD optimizer was used for model optimization. The model was trained with 4 GPUs on an NVIDIA DGX-1.

The trained query encoder of the MoCo model was used as the feature extractor to generate a 128×1 feature vector for each tile. Feature vectors of tiles from each slide were combined with aggregation models, which were finetuned with progression labels. We used a learning rate of 0.001 and a batch size of 64 for Deep-Att-MIL model training. The learning rate and batch size for Deep-Att-MIL-clinical were set at 0.001 and 12, respectively. Adam optimizer was used in both experiments. The learning rate was decreased 5 times if the validation loss didn't decrease for 2 epochs. Deep-GCN and Deep-GCN + clinical models were trained with a batch size of 16 and a learning rate of 0.0005. Adam optimizer was also used for parameter optimization for GCN models. Graphs were constructed by connecting 3 nearest neighbors for each node (*i.e., tile*).

6.3.2 Evaluation metrics

PFS prediction models are evaluated by the concordance index (c-index), also referred as c-statistic. c-index can be considered as a generalized area under receiver operating characteristic curve (AUROC). It measures the effectiveness of predicted risk scores (*i.e.,* hazard) on ranking survival times (*e.g.,* time to progression). The c-index value is between 0 and 1, with 1 indicating the perfect concordance and 0.5 representing results from random predictions. To compute c-index, slide-level predicted hazards are fed into the concordance_index function in the lifelines package [DKJ20]. For cases with multiple slides, average hazards are utilized.

6.3.3 Comparison models

Deep-Att-MIL denotes the deep survival model that leverages self-supervised learning features and attention MIl-based aggregation. Topological structures of learned deep representations are exploited in the GCN-based aggregation model, referred as *Deep-GCN*. Moreover, in *Deep-Att-MIL + clinical* and *Deep-GCN + clinical* we incorporate clinical features by concatenating clinical variables with aggregated slide-level feature vector, which is then utilized for survival estimation.

To further evaluate the effectiveness of deep representations and aggregation methods, we implemented several baseline models. Mean pooling was used to combine tile-level information for both handcrafted feature-based models and deep feature-based models.

Clinical features. In this experiment, we utilize clinical variables including age, psa value, pathological T stage and GG to fit a baseline Cox model. GG and age information are available for all cases. Missing values of psa and pathological T stage are imputed with median.

LASSO-Cox. Linear Cox models with $\mathcal{L}_1$ penalty are implemented for PFS prediction. Handcrafted features as described in §6.1.2.2 are used as covariates in the *LASSO-Cox-*

Table 6.2: Concordance index (c-index) of models using different feature pipelines and aggregation methods on predicting progression-free survival for prostate patients after radical prostatectomy.

Models	c-index ↑
Clinical features	0.7254 ± 0.042
LASSO-Cox-handcrafted	0.6818 ± 0.067
LASSO-Cox-deep	0.7046 ± 0.111
ElasticNet-Cox-handcrafted	0.6671 ± 0.052
ElasticNet-Cox-deep	0.7254 ± 0.101
RandomForest-Cox-handcrafted	0.6995 ± 0.048
RandomForest-Cox-deep	0.7209 ± 0.109
Deep-Att-MIL	0.7420 ± 0.089
Deep-GCN	**0.7441 ± 0.123**
RandomForest-Cox-handcrafted + clinical	0.7058 ± 0.052
ElasticNet-Cox-deep + clinical	0.7312 ± 0.049
Deep-Att-MIL + clinical	0.7602 ± 0.073
Deep-GCN + clinical	**0.7743 ± 0.122**

handcrafted experiment. The *LASSO-Cox-deep* experiment, on the other hand, leverages self-supervised learning features.

ElasticNet-Cox. We implement Cox models with $\mathcal{L}_1$ and $\mathcal{L}_2$ penalties in this experiment. The ratio of $\mathcal{L}_1$ penalty is set to be 0.5. The *ElasticNet-Cox-deep* utilizes deep representations, while the *ElasticNet-Cox-handcrafted* relies on handcrafted features.

RandomForest-Cox. The random forest-based Cox model is used for survival prediction. In *RandomForest-Cox-handcrafted*, handcrafted features are utilized. Self-supervised learning based representations are used as predictors in *RandomForest-Cox-deep*.

6.4 Results

Hazard predictions on the left-out test set for each fold are collected and utilized to compute c-index. Average c-index and 95% confidence interval for models and ablation experiments are shown in the Table 6.2. The baseline Cox model with clinical features (age, psa, pathological T stage and GG) obtained an average c-index of 72.54%. The Deep-GCN model that combines clinical variables achieved the highest average c-index of 77.43 %, which is over 5% higher than the baseline model with only clinical features.

Models with deep learning representations demonstrated superior performances comparing with handcrafted features. For instance, the ElasticNet-Cox model with deep representations produced an average c-index of 72.54%, which is around 6% higher than the one with handcrafted features. GCN based aggregation method showed better results in predicting progression comparing with ones with the att-MIL aggregation. The Deep-GCN + clinical, for example, achieved around 1% higher average c-index than the Deep-Att-MIL + clinical. Both Deep-GCN and Deep-Att-MIL presented better performances when fused feature vectors with clinical information. For instance, the Deep-Att-MIL with clinical data obtained around 2% higher average c-index than the one without clinical features.

6.5 Discussion

The self-supervised learning-based method is able to exploit intrinsic structures of images and doesn't rely on task specific labels. As demonstrated by ablation experiments, average c-indexs for several Cox models were largely improved by using features produced from self-supervised learning models.

Given the enormous size of WSIs and memory capacity of current GPUs, it's almost impossible to feed the entire WSI into a CNN model. Therefore, dividing the large slide into multiple smaller tiles is one potential way to reduce computational requirement and retain sufficient resolution. Representations are then extracted from each tiles. Attention MIL, which is invariant to tile permutation, is one of the most commonly used strategy for combining local features into global representations. However, this method discards spatial orders of tiles. Unlike the attention mechanism, GCN considers location information of tile-level features and updates each node feature with representations from nearby nodes. This enables learning of longer range dependency and facilitates spatial-aware aggregation of local features. As demonstrated in Table 6.2, models (*i.e.,* Deep-GCN and Deep-GCN + clinical) with GCN based aggregation generally outperformed attention-based models (*i.e.,* Deep-Att-MIL and Deep-Att-MIL + clinical).

Clinical factors are good predictors for cancer progression prediction and the baseline Cox model achieved an average c-index of 72.54%. Yet, current grading system describes histological growth patterns with fixed categories, which may not be able to account for the underlying diverse patterns. As shown in Table 6.2, models with deep representations and effective aggregation methods generally outperformed the baseline model with only clinical features, which indicates the ability of self-supervised learning models in identifying informative histopathological features for progression prediction. Performances of both deep representation-based models and handcrafted feature-based models improved by incorporating clinical features. The best performance is achieved by aggregating deep features with GCN and combining clinical variables with deep representations. This demonstrates the potential of including deep features for prostate cancer progression prediction.

Though the proposed deep learning system achieves promising performances on progression prediction, there are several limitations of this work. Different treatments and surgical margins could be confounding factors for progression and may affect model performances. Effects of treatment options and residual tumor should be investigated in the future work. Models were only evaluated retrospectively with 399 cases from the TCGA-PRAD dataset. External validation on larger datasets and prospective studies could be performed in the future work to further evaluate model performances. Moreover, in this study, we mainly leverage information embedded in histopathological images and clinical measurements to estimate progression free survival. Representations from other data modalities, such as multiparametric-magnetic resonance imaging (mp-MRI) and genomic profiling, may also contain important information about prostate cancer prognosis. Multi-modal models that incorporate features from different scales can be investigated in the future work.

6.6 Summary

In this Chapter, we present a deep learning system for progression-free survival prediction of prostate cancer patients. The system leverages histopathological features embedded in high-resolution histopathological tiles and effective aggregation approaches for progression estimation. Models are developed on the publicly available TCGA-PRAD dataset and evaluated on hold-out sets in 5-fold cross validation. Deep representations produced by the self-supervised learning model demonstrate superior performances comparing with handcrafted features. By representing tiles as nodes in the graph, the GCN-based aggregation method is able to model spatial distributions of tile-level features and shows better performance on predicting prostate cancer progression than using a permutation invariant attention module. In addition, models with deep representations outperform the baseline model with clinical feature only. The best model, which uses deep features, exploits GCN aggregation and combines clinical information, achieves an average c-index of 77.26% (around 5% higher than the baseline model). Though other factors such as treatment need to be included and the model performance needs to be further validated on large multi-institutional datasets, current results suggest the potential of leveraging quantitative histopathological features for better progression prediction.

6.7 Appendix

6.7.1 Tumor detection stage model performance

Tumor detection stage model is evaluated with the average precision (AP) and AUROC, which are computed using the scikit-learn package [PVG11].

The weakly-supervised cancer classification model, which is trained on the dataset-biopsy with only slide-level labels, achieves an AP of 0.9708 and an AUROC of 0.9842 for classifying benign and malignant slides on an independent test set. The model is further finetuned on a radical prostatectomy dataset with tile-level labels retrieved from region-level annotations. It achieves the AP of 0.9833 ± 0.010 and the AUROC of 0.9788 ± 0.012 for tile-level cancer prediction on 5-fold cross validation. This demonstrates the ability of the tumor detection stage on identifying suspicious regions. Figure 6.2 shows some examples of cancer probability maps.

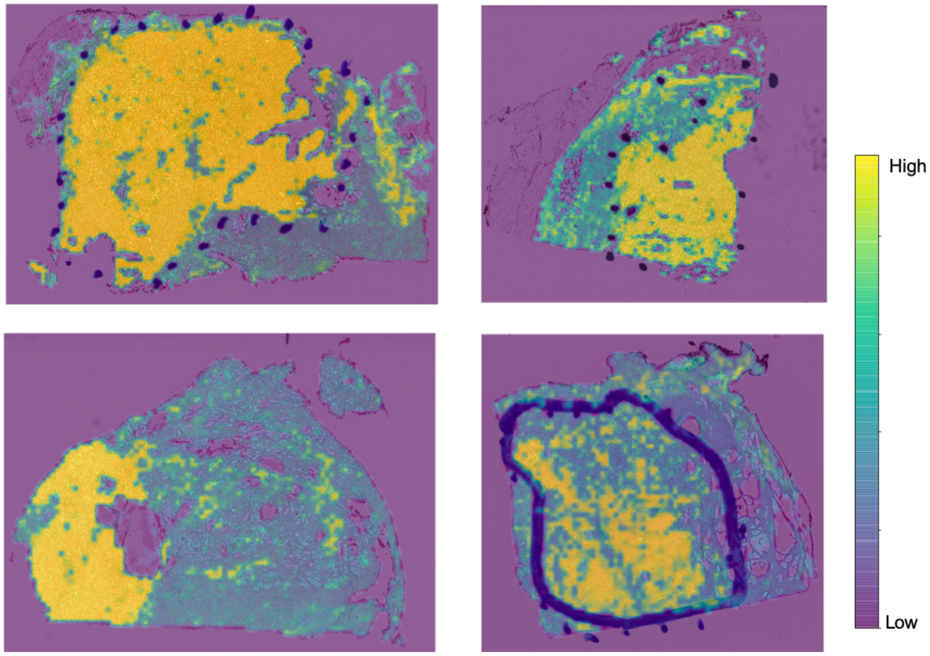

Figure 6.2: Examples of cancer probability maps produced from the tile-level cancer classification model. Top tiles with highest cancer probabilities are selected for PFS prediction.

CHAPTER 7

Conclusion and Future Work

This Chapter summarizes main contributions of this dissertation and discusses potential future directions to investigate.

7.1 Summary of contributions

This dissertation presents models for learning discriminative representations from large-scale whole slide images, which can potentially be incorporated into the pathology workflow to improve the efficiency and reproducibility of prostate cancer diagnosis. Contributions of this dissertation are summarized as follows.

1. A multi-scale U-Net model, which extends the U-Net model by explicitly leveraging features from different scales, is developed for semantic segmentation of heterogeneous histopathological images extracted from radical prostatectomy slides. As presented in Chapter 3, the proposed multi-scale U-Net model outperforms a reference segmentation algorithm based on handcrafted features, the pixel-wise deep CNN model and the original U-Net model. The segmentation model enables localization of tissues of different types and may facilitate estimation of Gleason pattern percentage.

2. An EM-based model that addresses the challenge of lack of data with fined-grained annotations for segmentation models is built to further improve the performance of the multi-scale U-Net model. As demonstrated in Chapter 4, the EM-based method leverages information embedded in the weakly-labeled dataset and exploits the prior

knowledge to refine prediction masks. The EM method with adaptive bias based on the prior knowledge shows superior performance than the baseline EN model and the one with fixed bias. With the proposed algorithm, segmentation models can be fine-tuned on the dataset with only image-level labels, which greatly reduces the need of costly pixel-wise annotations from pathologists.

3. A multi-resolution multiple instance learning-based (MRMIL) model is developed to address the challenge of training a classification model with large-scale weakly-labeled whole slide images. The proposed model could produce slide-level Gleason grade prediction as well as localize suspicious areas. Visualization techniques as t-SNE [MH08] and Grad-CAM [SCD17] are exploited to better understand the learned model. The MRMIL outperforms baseline models and achieves significant performances especially on differentiating malignant and benign slides.

4. The effectiveness of self-supervised learning-based features and different aggregation methods on characterizing underlying histopathological patterns is investigated in a progression-free survival prediction model. The proposed progression prediction system consists of two main stages: the tumor detection stage and the progression-free survival estimation stage with discriminative features extracted and aggregated from tiles. Deep histopathological features demonstrate superior performances comparing with models using handcrafted features and clinical factors in estimating progression. Better performances are achieved by combining clinical variables with deep representations. This study shows the potential of utilizing quantitative deep features for better progression prediction.

7.2 Future work

The computer aided diagnosis (CAD) tool for whole slide image classification and detection (detailed in Chapter 5) has been developed on a large-scale prostate biopsy dataset containing over 14k slides, and validated on an independent internal test set as well as an external dataset. Yet, validation with data from multiple institutions may be needed to further evaluate the generalizability of the model. Though promising results have been obtained for the MRMIL model, additional clinical studies with more evaluation metrics need to be performed to measure the efficacy of incorporating a CAD tool in current pathology workflow. User studies may need to be conducted to investigate effective ways to present model outputs to pathologists and to enable better interactions with the CAD tool. CAD tools are not designed to replace pathologists, but developed to assist pathologists during diagnosis. Feedback provided by pathologists when utilizing CAD tools could help improve performances of models. Training a new model every time to incorporate new feedback may be computationally infeasible. Online machine learning algorithms, which could continuously optimize models with signals from pathologists, can be an interesting direction to investigate in the future. Moreover, deep ensemble models may be investigated to improve model performances.

Interpretability of deep learning models is essential for building reliable, trustworthy and transparent diagnosis tools that could be adopted by pathologists in routine evaluation of histopathological slides. This dissertation mainly exploits methods to explain decision made by deep learning systems and to understand structures of learned features. For instance, attention maps, which indicate tiles that contribute most to the model prediction, are qualitatively evaluated by comparing with pen makers left by pathologists during diagnosis. Grad-CAM [SCD17] then identify informative regions within a tile. t-SNE [MH08], on the other hand, visualize lower dimensional manifolds of learned representations. However, how to understand biological meaning of learned deep features and how to build intrinsically explainable models that leverage domain specific knowledge such as causality, physical constraints and *etc* remain challenges[Rud19]. Models that are designed according to domain specific knowledge and constraints may be studied in the future work. Besides providing explanations of deep learning models, how to leverage interpretations to shed light on model design and improvement may need to be further investigated.

This dissertation investigates the effectiveness of quantitative histopathological features in predicting progression-free survival in prostate cancer patients. Yet, besides PSA, age, Gleason grade group and pathological T stage, effects of different treatment options pre- and after radical prostatectomy, and surgical margins, which could also affect the prostate cancer progression, should be considered in the future work.

Quality control and stain normalization are also important preprocessing steps to perform before feeding data into deep or machine learning systems. This dissertation utilizes intensity threshold-based method to remove noisy background and mainly exploits a computationally efficient stain normalization method [RAG01], which is based on color transformation. Image to image translation models based on generative adversarial network have been proposed for more robust stain normalization [ZCH19, NCS20]. These models can be investigated and incorporated into the current preprocessing pipeline to improve performances on downstream prediction tasks. Though with large amount of data, deep learning models can be robust to small percent of noisy samples, automated detection and removal of low quality slides (*e.g.,* slides with out of focus regions, extensive artifacts and *etc*) could help construct high quality datasets for research and clinical uses in the future.

Histopathological evaluation serves as a bridge between radiology and genomics. Morphological features of histologic primitives (*e.g.,* glands and nuclei) may reflect the underlying changes in molecular pathways. Alterations of tissue structures may be observed as suspicious lesions on medical imaging, which could produce an overall view of tumor appearances. Pathology, radiology and genomics provide unique pieces of information about tumor characteristics from different scales. In this dissertation, a deep survival model, which leverages deep representations from whole slide images and clinical information, has been developed. Yet, multi-modal hierarchical models, which can integrate the spectrum of a patient's diagnostic data including clinical information, imaging data, histopathological slides and molecular data to predict important prostate cancer endpoints, should be investigated to better guide clinical decision making in the future work.

Whole slide images are cross sections of complex 3D tissue structures. Artifacts may occur during the 2D projection such as tangential cut. 3D reconstruction with serially cuts tissue slides can provide better visualization of tumor micro-environment and facilitate understanding of diverse growth patterns [KBG20, STB13]. Registration algorithms for 3D reconstruction of whole slide images and representations learning with volumetric slides can be another interesting direction to pursue in the future.

In addition, this dissertation mainly focuses on development and validation of deep learning models for prostate cancer diagnosis. Algorithms and pipelines developed in this dissertation could be transferred to other disease domain in next steps. Pan-cancer models, which uncovers deep representations that are informative across different disease types, could also be investigated in the future work.

REFERENCES

[AAA15] Adam Abeshouse, Jaeil Ahn, Rehan Akbani, Adrian Ally, Samirkumar Amin, Christopher D Andry, Matti Annala, Armen Aprikian, Joshua Armenia, Arshi Arora, et al. "The molecular taxonomy of primary prostate cancer." *Cell*, **163**(4):1011–1025, 2015.

[AAB16] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. "Tensorflow: Large-scale machine learning on heterogeneous distributed systems." *arXiv preprint arXiv:1603.04467*, 2016.

[AC18] Eirini Arvaniti and Manfred Claassen. "Coupling weak and strong supervision for classification of prostate cancer histopathology images." *arXiv preprint arXiv:1811.07013*, 2018.

[Aga18] Abien Fred Agarap. "Deep learning using rectified linear units (relu)." *arXiv preprint arXiv:1803.08375*, 2018.

[AMJ01] William C Allsbrook Jr, Kathy A Mangold, Maribeth H Johnson, Roger B Lane, Cynthia G Lane, and Jonathan I Epstein. "Interobserver reproducibility of Gleason grading of prostatic carcinoma: general pathologist." *Human pathology*, **32**(1):81–88, 2001.

[Amo13] Jaume Amores. "Multiple instance classification: Review, taxonomy and comparative study." *Artificial intelligence*, **201**:81–105, 2013.

[ATH02] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. "Support vector machines for multiple-instance learning." *Advances in neural information processing systems*, **15**:577–584, 2002.

[ATH03] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. "Support vector machines for multiple-instance learning." In *Advances in neural information processing systems*, pp. 577–584, 2003.

[AYT18] Md Zahangir Alom, Chris Yakopcic, Tarek M Taha, and Vijayan K Asari. "Nuclei segmentation with recurrent residual convolutional neural networks based U-Net (R2U-Net)." In *NAECON 2018-IEEE National Aerospace and Electronics Conference*, pp. 228–233. IEEE, 2018.

[BHB19] Philip Bachman, R Devon Hjelm, and William Buchwalter. "Learning representations by maximizing mutual information across views." In *Advances in Neural Information Processing Systems*, pp. 15535–15545, 2019.

[Bis06] Christopher M Bishop. *Pattern recognition and machine learning.* springer, 2006.

[BKC17] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." *IEEE transactions on pattern analysis and machine intelligence*, **39**(12):2481–2495, 2017.

[BME13] Fadi Brimo, Rodolfo Montironi, Lars Egevad, Andreas Erbersdobler, Daniel W Lin, Joel B Nelson, Mark A Rubin, Theo Van Der Kwast, Mahul Amin, and Jonathan I Epstein. "Contemporary grading for prostate cancer: implications for patient care." *European urology*, **63**(5):892–901, 2013.

[BPB19] Wouter Bulten, Hans Pinckaers, Hester van Boven, Robert Vink, Thomas de Bel, Bram van Ginneken, Jeroen van der Laak, Christina Hulsbergen-van de Kaa, and Geert Litjens. "Automated gleason grading of prostate biopsies using deep learning." *arXiv preprint arXiv:1907.07980*, 2019.

[BPB20] Wouter Bulten, Hans Pinckaers, Hester van Boven, Robert Vink, Thomas de Bel, Bram van Ginneken, Jeroen van der Laak, Christina Hulsbergen-van de Kaa, and Geert Litjens. "Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study." *The Lancet Oncology*, **21**(2):233–241, 2020.

[BSR19] Kaustav Bera, Kurt A Schalper, David L Rimm, Vamsidhar Velcheti, and Anant Madabhushi. "Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology." *Nature reviews Clinical oncology*, **16**(11):703–715, 2019.

[BVV17] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer." *Jama*, **318**(22):2199–2210, 2017.

[CBP19] Veronika Cheplygina, Marleen de Bruijne, and Josien PW Pluim. "Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis." *Medical image analysis*, **54**:280–296, 2019.

[CBW06] Yixin Chen, Jinbo Bi, and James Ze Wang. "MILES: Multiple-instance learning via embedded instance selection." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**(12):1931–1947, 2006.

[CCG18] Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. "Multiple instance learning: A survey of problem characteristics and applications." *Pattern Recognition*, **77**:329–353, 2018.

[CDL07] Liang Cheng, Darrell D Davidson, Haiqun Lin, and Michael O Koch. "Percentage of Gleason pattern 4 and 5 predicts survival after radical prostatectomy." *Cancer*, **110**(9):1967–1972, 2007.

[CDS18] Lee Ad Cooper, Elizabeth G Demicco, Joel H Saltz, Reid T Powell, Arvind Rao, and Alexander J Lazar. "PanCancer insights from The Cancer Genome Atlas: the pathologist's perspective." *The Journal of pathology*, **244**(5):512–524, 2018.

[CG16] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system." In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[CGB17] Francesco Ciompi, Oscar Geessink, Babak Ehteshami Bejnordi, Gabriel Silva De Souza, Alexi Baidoshvili, Geert Litjens, Bram Van Ginneken, Iris Nagtegaal, and Jeroen Van Der Laak. "The importance of stain normalization in colorectal tissue classification with convolutional networks." In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pp. 160–163. IEEE, 2017.

[CGG12] Dan Ciresan, Alessandro Giusti, Luca Gambardella, and Jürgen Schmidhuber. "Deep neural networks segment neuronal membranes in electron microscopy images." *Advances in neural information processing systems*, **25**:2843–2851, 2012.

[CGZ20] Lei Cai, Jingyang Gao, and Di Zhao. "A review of the application of deep learning in medical image classification and segmentation." *Annals of translational medicine*, **8**(11), 2020.

[CHG19] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images." *Nature medicine*, **25**(8):1301–1309, 2019.

[Cho18] François Chollet et al. "Keras: The python deep learning library." *ascl*, pp. ascl–1806, 2018.

[CKF11] Ronan Collobert, Koray Kavukcuoglu, and Clément Farabet. "Torch7: A matlab-like environment for machine learning." In *BigLearn, NIPS workshop*, number CONF, 2011.

[CKN20] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. "A simple framework for contrastive learning of visual representations." *arXiv preprint arXiv:2002.05709*, 2020.

[CLL20] Sacheth Chandramouli, Patrick Leo, George Lee, Robin Elliott, Christine Davis, Guangjing Zhu, Pingfu Fu, Jonathan I Epstein, Robert Veltri, and Anant Madabhushi. "Computer Extracted Features from Initial H&E Tissue Biopsies Predict Disease Progression for Prostate Cancer Patients on Active Surveillance." *Cancers*, **12**(9):2708, 2020.

[CLP12] Hang Chang, Leandro A Loss, and Bahram Parvin. "Nuclear segmentation in H&E sections via multi-reference graph cut (MRGC)." In *International symposium biomedical imaging*, 2012.

[CLP13] Gabriela Csurka, Diane Larlus, Florent Perronnin, and France Meylan. "What is a good evaluation measure for semantic segmentation?." In *BMVC*, volume 27, p. 2013, 2013.

[CMS16] Adam I Cole, Todd M Morgan, Daniel E Spratt, Ganesh S Palapattu, Chang He, Scott A Tomlins, Alon Z Weizer, Felix Y Feng, Angela Wu, Javed Siddiqui, et al. "Prognostic value of percent Gleason grade 4 at prostate biopsy in predicting prostatectomy pathology and recurrence." *The Journal of urology*, **196**(2):405–411, 2016.

[Cox72] David R Cox. "Regression models and life-tables." *Journal of the Royal Statistical Society: Series B (Methodological)*, **34**(2):187–202, 1972.

[CPK17] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." *IEEE transactions on pattern analysis and machine intelligence*, **40**(4):834–848, 2017.

[CRH19] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. "Human-centered tools for coping with imperfect algorithms during medical decision-making." In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2019.

[CTL15] Veronika Cheplygina, David MJ Tax, and Marco Loog. "Multiple instance learning with bag dissimilarities." *Pattern recognition*, **48**(1):264–275, 2015.

[DAH18] Neofytos Dimitriou, Ognjen Arandjelović, David J Harrison, and Peter D Caie. "A principled machine learning framework improves accuracy of stage II colorectal cancer prognosis." *npj Digital Medicine*, **1**(1):1–9, 2018.

[DDS09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database." In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. IEEE, 2009.

[DFT10] Scott Doyle, Michael Feldman, John Tomaszewski, and Anant Madabhushi. "A boosted Bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies." *IEEE transactions on biomedical engineering*, **59**(5):1205–1218, 2010.

[DKJ20] Cameron Davidson-Pilon, Jonas Kalderstam, Noah Jacobson, Ben Kuhn, Paul Zivich, Mike Williamson, Deepyaman Datta, Andrew Fiore-Gartland, Alex Parij, Daniel WIlson, et al. "CamDavidsonPilon/lifelines: v0. 24.15." *Zenodo*, 2020.

[DLL97] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. "Solving the multiple instance problem with axis-parallel rectangles." *Artificial intelligence*, **89**(1-2):31–71, 1997.

[DTC16] Thibaut Durand, Nicolas Thome, and Matthieu Cord. "Weldon: Weakly supervised learning of deep convolutional neural networks." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4743–4752, 2016.

[DZY20] Shujian Deng, Xin Zhang, Wen Yan, I Eric, Chao Chang, Yubo Fan, Maode Lai, and Yan Xu. "Deep learning in digital pathology image analysis: a survey." *Frontiers of Medicine*, pp. 1–18, 2020.

[EEV15] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. "The pascal visual object classes challenge: A retrospective." *International journal of computer vision*, **111**(1):98–136, 2015.

[ELC19] Ángel E Esteban, Miguel López-Pérez, Adrián Colomer, María A Sales, Rafael Molina, and Valery Naranjo. "A new optical density granulometry-based descriptor for the classification of prostate histological images using shallow and deep Gaussian processes." *Computer methods and programs in biomedicine*, **178**:303–317, 2019.

[EVW10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. "The pascal visual object classes (voc) challenge." *International journal of computer vision*, **88**(2):303–338, 2010.

[EZS16] Jonathan I Epstein, Michael J Zelefsky, Daniel D Sjoberg, Joel B Nelson, Lars Egevad, Cristina Magi-Galluzzi, Andrew J Vickers, Anil V Parwani, Victor E Reuter, Samson W Fine, et al. "A contemporary prostate cancer grading system: a validated alternative to the Gleason score." *European urology*, **69**(3):428–435, 2016.

[FAB12] Samson W Fine, Mahul B Amin, Daniel M Berney, Anders Bjartell, Lars Egevad, Jonathan I Epstein, Peter A Humphrey, Christina Magi-Galluzzi, Rodolfo Montironi, and Christian Stief. "A contemporary update on pathology reporting for prostate cancer: biopsy and radical prostatectomy specimens." *European urology*, **62**(1):20–39, 2012.

[FM82] Kunihiko Fukushima and Sei Miyake. "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition." In *Competition and cooperation in neural nets*, pp. 267–285. Springer, 1982.

[FS95] David Faraggi and Richard Simon. “A neural network model for survival data.” *Statistics in medicine*, **14**(1):73–82, 1995.

[FSJ07] Reza Farjam, Hamid Soltanian-Zadeh, Kourosh Jafari-Khouzani, and Reza A Zoroofi. “An image analysis approach for automatic malignancy determination of prostate pathological images.” *Cytometry Part B: Clinical Cytometry: The Journal of the International Society for Analytical Cytology*, **72**(4):227–240, 2007.

[FWD18] Joshua J Fenton, Meghan S Weyrich, Shauna Durbin, Yu Liu, Heejung Bang, and Joy Melnikow. “Prostate-specific antigen–based screening for prostate cancer: evidence report and systematic review for the US Preventive Services Task Force.” *Jama*, **319**(18):1914–1931, 2018.

[FZ17] Ji Feng and Zhi-Hua Zhou. “Deep MIML Network.” In *AAAI*, pp. 1884–1890, 2017.

[Gar16] Marcial García-Rojo. “International clinical guidelines for the adoption of digital pathology: a review of technical aspects.” *Pathobiology*, **83**(2-3):99–109, 2016.

[GBO08] Elisa Drelie Gelasca, Jiyun Byun, Boguslaw Obara, and BS Manjunath. “Evaluation and benchmark for biological image segmentation.” In *2008 15th IEEE International Conference on Image Processing*, pp. 1816–1819. IEEE, 2008.

[GC11] Maya R Gupta and Yihua Chen. *Theory and use of the EM algorithm.* Now Publishers Inc, 2011.

[GGA14] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. “Learning rich features from RGB-D images for object detection and segmentation.” In *European conference on computer vision*, pp. 345–360. Springer, 2014.

[GIM15] Arkadiusz Gertych, Nathan Ing, Zhaoxuan Ma, Thomas J Fuchs, Sadri Salman, Sambit Mohanty, Sanica Bhele, Adriana Velásquez-Vacca, Mahul B Amin, and Beatrice S Knudsen. “Machine learning approaches to analyze histological images of tissues from radical prostatectomies.” *Computerized Medical Imaging and Graphics*, **46**:197–208, 2015.

[GMS05] Marco Gori, Gabriele Monfardini, and Franco Scarselli. “A new model for learning in graph domains.” In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pp. 729–734. IEEE, 2005.

[GVG13] Lena Gorelick, Olga Veksler, Mena Gaed, José A Gómez, Madeleine Moussa, Glenn Bauman, Aaron Fenster, and Aaron D Ward. “Prostate histopathology: Learning tissue component histograms for cancer detection and classification.” *IEEE transactions on medical imaging*, **32**(10):1804–1818, 2013.

[HAG14] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. "Simultaneous detection and segmentation." In *European Conference on Computer Vision*, pp. 297–312. Springer, 2014.

[HDL16] Freddie C Hamdy, Jenny L Donovan, J Lane, Malcolm Mason, Chris Metcalfe, Peter Holding, Michael Davis, Tim J Peters, Emma L Turner, Richard M Martin, et al. "10-year outcomes after monitoring, surgery, or radiotherapy for localized prostate cancer." *N Engl J Med*, **375**:1415–1424, 2016.

[HDW17] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. "Brain tumor segmentation with deep neural networks." *Medical image analysis*, **35**:18–31, 2017.

[HFW20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. "Momentum contrast for unsupervised visual representation learning." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

[HGD17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. "Mask r-cnn." In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

[HHL19] Narayan Hegde, Jason D Hipp, Yun Liu, Michael Emmert-Buck, Emily Reif, Daniel Smilkov, Michael Terry, Carrie J Cai, Mahul B Amin, Craig H Mermel, et al. "Similar image search for histopathology: SMILY." *NPJ digital medicine*, **2**(1):1–9, 2019.

[HJW18] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. "Self-Erasing Network for Integral Object Attention." In *Advances in Neural Information Processing Systems*, pp. 549–559, 2018.

[HKZ14] Cheng Cheng Huang, Max Xiangtian Kong, Ming Zhou, Andrew B Rosenkrantz, Samir S Taneja, Jonathan Melamed, and Fang-Ming Deng. "Gleason score 3+ 4= 7 prostate cancer with minimal quantity of gleason pattern 4 on needle biopsy is associated with low-risk tumor in radical prostatectomy specimen." *The American journal of surgical pathology*, **38**(8):1096–1101, 2014.

[HLV17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. "Densely connected convolutional networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[HPE15] Matthew G Hanna, Liron Pantanowitz, and Andrew J Evans. "Overview of contemporary guidelines in digital pathology: what is available in 2015 and what still needs to be addressed?", 2015.

[HRS19] Matthew G Hanna, Victor E Reuter, Jennifer Samboy, Christine England, Lorraine Corsale, Samson W Fine, Narasimhan P Agaram, Evangelos Stamelos, Yukako Yagi, Meera Hameed, et al. "Implementation of digital pathology offers clinical and operational increase in efficiency and cost savings." *Archives of pathology & laboratory medicine*, **143**(12):1545–1555, 2019.

[HS13] Yonatan Halpern and David Sontag. "Unsupervised learning of noisy-or bayesian networks." *arXiv preprint arXiv:1309.6834*, 2013.

[HSK16] Le Hou, Dimitris Samaras, Tahsin M Kurc, Yi Gao, James E Davis, and Joel H Saltz. "Patch-based convolutional neural network for whole slide tissue image classification." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2424–2433, 2016.

[HUE19a] Achim Hekler, Jochen S Utikal, Alexander H Enk, Wiebke Solass, Max Schmitt, Joachim Klode, Dirk Schadendorf, Wiebke Sondermann, Cindy Franklin, Felix Bestvater, et al. "Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images." *European Journal of Cancer*, **118**:91–96, 2019.

[HUE19b] Achim Hekler, Jochen Sven Utikal, Alexander H Enk, Carola Berking, Joachim Klode, Dirk Schadendorf, Philipp Jansen, Cindy Franklin, Tim Holland-Letz, Dieter Krahl, et al. "Pathologist-level classification of histopathological melanoma images with deep neural networks." *European Journal of Cancer*, **115**:79–83, 2019.

[HZC17] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." *arXiv preprint arXiv:1704.04861*, 2017.

[HZR16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[IML18] Nathan Ing, Zhaoxuan Ma, Jiayun Li, Hootan Salemi, Corey Arnold, Beatrice S Knudsen, and Arkadiusz Gertych. "Semantic segmentation for prostate cancer grading by convolutional neural networks." In *Medical Imaging 2018: Digital Pathology*, volume 10581, p. 105811B. International Society for Optics and Photonics, 2018.

[IS15] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." *arXiv preprint arXiv:1502.03167*, 2015.

[ITW18] Maximilian Ilse, Jakub M Tomczak, and Max Welling. "Attention-based deep multiple instance learning." *arXiv preprint arXiv:1802.04712*, 2018.

[JHE17] Zhipeng Jia, Xingyi Huang, I Eric, Chao Chang, and Yan Xu. "Constrained deep weak supervision for histopathology image segmentation." *IEEE transactions on medical imaging*, **36**(11):2376–2388, 2017.

[JSD14] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. "Caffe: Convolutional architecture for fast feature embedding." In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 675–678, 2014.

[KB14] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980*, 2014.

[KBF16] Oren Z Kraus, Jimmy Lei Ba, and Brendan J Frey. "Classifying and segmenting microscopy images with deep multiple instance learning." *Bioinformatics*, **32**(12):i52–i59, 2016.

[KBG20] Ashley Kiemen, Alicia M Braxton, Mia P Grahn, Kyu Sang Han, Jaanvi Mahesh Babu, Rebecca Reichel, Falone Amoa, Seung-Mo Hong, Toby C Cornish, Elizabeth D Thompson, et al. "In situ characterization of the 3D microanatomy of the pancreas and pancreatic cancer at single cell resolution." *bioRxiv*, 2020.

[KH09] Alex Krizhevsky, Geoffrey Hinton, et al. "Learning multiple layers of features from tiny images." 2009.

[KRL91] James D Keeler, David E Rumelhart, and Wee Kheng Leow. "Integrated segmentation and recognition of hand-printed numerals." In *Advances in neural information processing systems*, pp. 557–563, 1991.

[KSC16] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. "Deep survival: A deep cox proportional hazards network." *stat*, **1050**(2), 2016.

[KSH17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks." *Communications of the ACM*, **60**(6):84–90, 2017.

[KZL10] Laurence Klotz, Liying Zhang, Adam Lam, Robert Nam, Alexandre Mamedov, and Andrew Loblaw. "Clinical results of long-term follow-up of a large, active surveillance cohort with localized prostate cancer." *Journal of Clinical Oncology*, **28**(1):126–131, 2010.

[LBB98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE*, **86**(11):2278–2324, 1998.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." *nature*, **521**(7553):436–444, 2015.

[LBN14] Stacy Loeb, Marc A Bjurlin, Joseph Nicholson, Teuvo L Tammela, David F Penson, H Ballentine Carter, Peter Carroll, and Ruth Etzioni. "Overdiagnosis and overtreatment of prostate cancer." *European urology*, **65**(6):1046–1055, 2014.

[LD12] Hugh J Lavery and Michael J Droller. "Do Gleason patterns 3 and 4 prostate cancer represent separate disease states?" *The Journal of urology*, **188**(5):1667–1675, 2012.

[LIS14] Emilie Lalonde, Adrian S Ishkanian, Jenna Sykes, Michael Fraser, Helen Ross-Adams, Nicholas Erho, Mark J Dunning, Silvia Halim, Alastair D Lamb, Nathalie C Moon, et al. "Tumour genomic and microenvironmental heterogeneity for integrated prediction of 5-year biochemical recurrence of prostate cancer: a retrospective cohort study." *The lancet oncology*, **15**(13):1521–1532, 2014.

[LJE19] Patrick Leo, Andrew Janowczyk, Robin Elliott, Nafiseh Janaki, Rakesh Shiradkar, Xavier Farrè, Kosj Yamoah, Timothy Rebbeck, Natalie Shih, Francesca Khani, et al. "Computerized histomorphometric features of glandular architecture predict risk of biochemical recurrence following radical prostatectomy: A multisite study.", 2019.

[LLG19] Jiayun Li, Wenyuan Li, Arkadiusz Gertych, Beatrice S Knudsen, William Speier, and Corey W Arnold. "An attention-based multi-resolution model for prostate whole slide image classification and localization." *arXiv preprint arXiv:1905.13208*, 2019.

[LLH18] Jianfang Liu, Tara Lichtenberg, Katherine A Hoadley, Laila M Poisson, Alexander J Lazar, Andrew D Cherniack, Albert J Kovatich, Christopher C Benz, Douglas A Levine, Adrian V Lee, et al. "An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics." *Cell*, **173**(2):400–416, 2018.

[LLS18] Wenyuan Li, Jiayun Li, Karthik V Sarma, King Chung Ho, Shiwen Shen, Beatrice S Knudsen, Arkadiusz Gertych, and Corey W Arnold. "Path R-CNN for prostate cancer diagnosis and gleason grading of histological images." *IEEE transactions on medical imaging*, **38**(4):945–954, 2018.

[LLS20] Jiayun Li, Wenyuan Li, Anthony Sisk, Huihui Ye, W Dean Wallace, William Speier, and Corey W Arnold. "A Multi-resolution Model for Histopathology

Image Classification and Localization with Multiple Instance Learning." *arXiv preprint arXiv:2011.02679*, 2020.

[LRW18] Cheng Lu, David Romo-Bucheli, Xiangxue Wang, Andrew Janowczyk, Shridar Ganesan, Hannah Gilmore, David Rimm, and Anant Madabhushi. "Nuclear shape and orientation features from H&E images predict survival in early-stage estrogen receptor-positive breast cancers." *Laboratory Investigation*, **98**(11):1438–1448, 2018.

[LSD15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

[LSF19] Peter Lawson, Jordan Schupbach, Brittany Terese Fasy, and John W Sheppard. "Persistent homology for the automatic classification of prostate cancer aggressiveness in histopathology images." In *Medical Imaging 2019: Digital Pathology*, volume 10956, p. 109560G. International Society for Optics and Photonics, 2019.

[LSH17] Jiayun Li, Karthik V Sarma, King Chung Ho, Arkadiusz Gertych, Beatrice S Knudsen, and Corey W Arnold. "A multi-scale U-Net for semantic segmentation of histological images from radical prostatectomies." In *AMIA Annual Symposium Proceedings*, volume 2017, p. 1140. American Medical Informatics Association, 2017.

[LSH18] Jiayun Li, William Speier, King Chung Ho, Karthik V Sarma, Arkadiusz Gertych, Beatrice S Knudsen, and Corey W Arnold. "An EM-based semi-supervised deep learning approach for semantic segmentation of histopathological images from radical prostatectomies." *Computerized Medical Imaging and Graphics*, **69**:125–133, 2018.

[LST16] Geert Litjens, Clara I Sánchez, Nadya Timofeeva, Meyke Hermsen, Iris Nagtegaal, Iringo Kovacs, Christina Hulsbergen-Van De Kaa, Peter Bult, Bram Van Ginneken, and Jeroen Van Der Laak. "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis." *Scientific reports*, **6**:26286, 2016.

[LW02] Andy Liaw, Matthew Wiener, et al. "Classification and regression by randomForest." *R news*, **2**(3):18–22, 2002.

[MAM17] Caner Mercan, Selim Aksoy, Ezgi Mercan, Linda G Shapiro, Donald L Weaver, and Joann G Elmore. "Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images." *IEEE transactions on medical imaging*, **37**(1):316–325, 2017.

[MAV14] Clara Mosquera-Lopez, Sos Agaian, Alejandro Velez-Hoyos, and Ian Thompson. "Computer-aided prostate cancer diagnosis from digitized histopathology: a review on texture-based systems." *IEEE reviews in biomedical engineering*, **8**:98–113, 2014.

[MBB10] James Mohler, Robert R Bahnson, Barry Boston, J Erik Busby, Anthony D'Amico, James A Eastham, Charles A Enke, Daniel George, Eric Mark Horwitz, Robert P Huben, et al. "Prostate cancer." *Journal of the National Comprehensive Cancer Network*, **8**(2):162–200, 2010.

[MBP20] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. "Image segmentation using deep learning: A survey." *arXiv preprint arXiv:2001.05566*, 2020.

[MGM14] Jaime Melendez, Bram van Ginneken, Pragnya Maduskar, Rick HHM Philipsen, Klaus Reither, Marianne Breuninger, Ifedayo MO Adetifa, Rahmatulai Maane, Helen Ayles, and Clara I Sánchez. "A novel multiple-instance learning-based approach to computer-aided detection of tuberculosis on chest x-rays." *IEEE transactions on medical imaging*, **34**(1):179–192, 2014.

[MGM15] Jaime Melendez, Bram van Ginneken, Pragnya Maduskar, Rick HHM Philipsen, Helen Ayles, and Clara I Sánchez. "On combining multiple-instance learning and active learning for computer-aided detection of tuberculosis." *IEEE transactions on medical imaging*, **35**(4):1013–1024, 2015.

[MH08] Laurens van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." *Journal of machine learning research*, **9**(Nov):2579–2605, 2008.

[MIW19] Brian Miles, Michael Ittmann, Thomas Wheeler, Mohammad Sayeeduddin, Antonio Cubilla, David Rowley, Ping Bu, Yi Ding, Yan Gao, MinJae Lee, et al. "Moving Beyond Gleason Scoring." *Archives of pathology & laboratory medicine*, **143**(5):565–570, 2019.

[ML97] Oded Maron and Tomás Lozano-Pérez. "A framework for multiple-instance learning." *Advances in neural information processing systems*, **10**:570–576, 1997.

[MMB18] Sachin Mehta, Ezgi Mercan, Jamen Bartlett, Donald Weaver, Joann G Elmore, and Linda Shapiro. "Y-Net: joint segmentation and classification for diagnosis of breast biopsy images." In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 893–901. Springer, 2018.

[MNA16] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-net: Fully convolutional neural networks for volumetric medical image segmentation." In *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571. IEEE, 2016.

[MNM09] Marc Macenko, Marc Niethammer, James S Marron, David Borland, John T Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E Thomas. “A method for normalizing histology slides for quantitative analysis.” In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 1107–1110. IEEE, 2009.

[MWH16] Jesse K McKenney, Wei Wei, Sarah Hawley, Heidi Auman, Lisa F Newcomb, Hilary D Boyer, Ladan Fazli, Jeff Simko, Antonio Hurtado-Coll, Dean A Troyer, et al. “Histologic grading of prostatic adenocarcinoma can be further optimized.” *The American journal of surgical pathology*, **40**(11):1439–1456, 2016.

[NCJ20] Soojeong Nam, Yosep Chong, Chan Kwon Jung, Tae-Yeong Kwak, Ji Youl Lee, Jihwan Park, Mi Jung Rho, and Heounjeong Go. “Introduction to digital pathology and computer-aided pathology.” *Journal of Pathology and Translational Medicine*, **54**(2):125, 2020.

[NCS20] Harshal Nishar, Nikhil Chavanke, and Nitin Singhal. “Histopathological Stain Transfer Using Style Transfer Network with Adversarial Loss.” In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 330–340. Springer, 2020.

[NFL19] Kunal Nagpal, Davis Foote, Yun Liu, Po-Hsuan Cameron Chen, Ellery Wulczyn, Fraser Tan, Niels Olson, Jenny L Smith, Arash Mohtashamian, James H Wren, et al. “Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer.” *npj Digital Medicine*, **2**(1):48, 2019.

[NFT20] Kunal Nagpal, Davis Foote, Fraser Tan, Yun Liu, Po-Hsuan Cameron Chen, David F Steiner, Naren Manoj, Niels Olson, Jenny L Smith, Arash Mohtashamian, et al. “Development and Validation of a Deep Learning Algorithm for Gleason Grading of Prostate Cancer From Biopsy Specimens.” *JAMA oncology*, 2020.

[NH10] Vinod Nair and Geoffrey E Hinton. “Rectified linear units improve restricted boltzmann machines.” In *ICML*, 2010.

[NHH15] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. “Learning deconvolution network for semantic segmentation.” In *Proceedings of the IEEE international conference on computer vision*, pp. 1520–1528, 2015.

[NKG19] Guy Nir, Davood Karimi, S Larry Goldenberg, Ladan Fazli, Brian F Skinnider, Peyman Tavassoli, Dmitry Turbin, Carlos F Villamil, Gang Wang, Darby JS Thompson, et al. “Comparison of artificial intelligence techniques to evaluate performance of a classifier for automatic grading of prostate cancer from digitized histopathologic images.” *JAMA network open*, **2**(3):e190442–e190442, 2019.

[NSJ12] Kien Nguyen, Bikash Sabata, and Anil K Jain. "Prostate cancer grading: Gland segmentation and structural features." *Pattern Recognition Letters*, **33**(7):951–961, 2012.

[OLV18] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding." *arXiv preprint arXiv:1807.03748*, 2018.

[OSY96] SO Ozdamar, S Sarikaya, L Yildiz, MK Atilla, B Kandemir, and S Yildiz. "Intraobserver and interobserver reproducibility of WHO and Gleason histologic grading systems in prostatic adenocarcinomas." *International urology and nephrology*, **28**(1):73, 1996.

[PC15] Pedro O Pinheiro and Ronan Collobert. "From image-level to pixel-level labeling with convolutional networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1713–1721, 2015.

[PCM15] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation." In *Proceedings of the IEEE international conference on computer vision*, pp. 1742–1750, 2015.

[PGC17] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. "Automatic differentiation in pytorch." 2017.

[PMB12] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. "Understanding the exploding gradient problem." *CoRR, abs/1211.5063*, **2**:417, 2012.

[PPP15] Sanoj Punnen, Nicola Pavan, and Dipen J Parekh. "Finding the wolf in sheep's clothing: the 4Kscore is a novel blood test that can accurately identify the risk of aggressive prostate cancer." *Reviews in urology*, **17**(1):3, 2015.

[PSL14] Deepak Pathak, Evan Shelhamer, Jonathan Long, and Trevor Darrell. "Fully convolutional multi-class multiple instance learning." *arXiv preprint arXiv:1412.7144*, 2014.

[PVG11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, **12**:2825–2830, 2011.

[QBD13] Amir Qaseem, Michael J Barry, Thomas D Denberg, Douglas K Owens, and Paul Shekelle. "Screening for prostate cancer: a guidance statement from the Clinical Guidelines Committee of the American College of Physicians." *Annals of internal medicine*, **158**(10):761–769, 2013.

[QLC16] Gwenolé Quellec, Mathieu Lamard, Michel Cozic, Gouenou Coatrieux, and Guy Cazuguel. "Multiple-instance learning for anomaly detection in digital mammography." *IEEE transactions on medical imaging*, **35**(7):1604–1614, 2016.

[RAG01] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. "Color transfer between images." *IEEE Computer graphics and applications*, **21**(5):34–41, 2001.

[RD00] Jan Ramon and Luc De Raedt. "Multi instance neural networks." In *Proceedings of the ICML-2000 workshop on attribute-value and relational learning*, pp. 53–60, 2000.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

[RKG18] Jian Ren, Kubra Karagoz, Michael Gatza, David J Foran, and Xin Qi. "Differentiation among prostate cancer patients with Gleason score of 7 using histopathology whole-slide image and genomic data." In *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*, volume 10579, p. 1057904. International Society for Optics and Photonics, 2018.

[RPA13] Varun Rastogi, Naveen Puri, Swati Arora, Geetpriya Kaur, Lalita Yadav, and Rachna Sharma. "Artefacts: a diagnostic dilemma–a review." *Journal of clinical and diagnostic research: JCDR*, **7**(10):2408, 2013.

[RTG00] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. "The earth mover's distance as a metric for image retrieval." *International journal of computer vision*, **40**(2):99–121, 2000.

[Rud19] Cynthia Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nature Machine Intelligence*, **1**(5):206–215, 2019.

[SA11] Alex Shteynshlyuger and Gerald L Andriole. "Cost-effectiveness of prostate specific antigen screening in the United States: extrapolating from the European study of screening for prostate cancer." *The Journal of urology*, **185**(3):828–832, 2011.

[SA12] Gurdarshan S Sandhu and Gerald L Andriole. "Overdiagnosis of prostate cancer." *Journal of the National Cancer Institute Monographs*, **2012**(45):146–151, 2012.

[SAT16] Korsuk Sirinukunwattana, Shan e Ahmed Raza, Yee-Wah Tsang, David RJ Snead, Ian A Cree, and Nasir M Rajpoot. "Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images." *IEEE transactions on medical imaging*, **35**(5):1196–1206, 2016.

[SCD17] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." In *ICCV*, pp. 618–626, 2017.

[SCM20] Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. "Deep neural network models for computational histopathology: A survey." *Medical Image Analysis*, p. 101813, 2020.

[SCR18] Monjoy Saha, Chandan Chakraborty, and Daniel Racoceanu. "Efficient deep learning model for mitosis detection using breast histopathology images." *Computerized Medical Imaging and Graphics*, **64**:29–40, 2018.

[SFH11] Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. "Regularization paths for Cox's proportional hazards model via coordinate descent." *Journal of statistical software*, **39**(5):1, 2011.

[SHK14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting." *The journal of machine learning research*, **15**(1):1929–1958, 2014.

[SL17] Krishna Kumar Singh and Yong Jae Lee. "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization." In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3544–3553. IEEE, 2017.

[SLD17] Evan Shelhamer, Jonathan Long, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." *IEEE transactions on pattern analysis and machine intelligence*, **39**(4):640–651, 2017.

[SLJ15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[SM85] Hans Svanholm and Henrik Mygind. "Prostatic carcinoma reproducibility of histologic grading." *Acta Pathologica Microbiologica Scandinavica Series A: Pathology*, **93**(1-6):67–71, 1985.

[SMJ19] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. "Cancer statistics, 2019." *CA: a cancer journal for clinicians*, **69**(1):7–34, 2019.

[SML18] David F Steiner, Robert MacDonald, Yun Liu, Peter Truszkowski, Jason D Hipp, Christopher Gammage, Florence Thng, Lily Peng, and Martin C Stumpe. "Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer." *The American journal of surgical pathology*, **42**(12):1636, 2018.

[SPS19] PJ Sudharshan, Caroline Petitjean, Fabio Spanhol, Luiz Eduardo Oliveira, Laurent Heutte, and Paul Honeine. "Multiple instance learning for histopathological breast cancer image classification." *Expert Systems with Applications*, **117**:103–111, 2019.

[STB13] Yi Song, Darren Treanor, Andrew J Bulpitt, and Derek R Magee. "3D reconstruction of multiple stained histology images." *Journal of pathology informatics*, **4**(Suppl), 2013.

[SYS18] Krishna Kumar Singh, Hao Yu, Aron Sarmasi, Gautam Pradeep, and Yong Jae Lee. "Hide-and-Seek: A Data Augmentation Technique for Weakly-Supervised Localization and Beyond." *arXiv preprint arXiv:1811.02545*, 2018.

[SZ14] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556*, 2014.

[TAO17] Oscar Jiménez del Toro, Manfredo Atzori, Sebastian Otálora, Mats Andersson, Kristian Eurén, Martin Hedlund, Peter Rönnquist, and Henning Müller. "Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade gleason score." In *Medical Imaging 2017: Digital Pathology*, volume 10140, p. 101400O. International Society for Optics and Photonics, 2017.

[TBO19] Ruwan Tennakoon, Gerda Bortsova, Silas Ørting, Amirali K Gostar, Mathilde MW Wille, Zaigham Saghir, Reza Hoseinnezhad, Marleen de Bruijne, and Alireza Bab-Hadiashar. "Classification of Volumetric Images Using Multi-Instance Learning and Extreme Value Theorem." *IEEE transactions on medical imaging*, 2019.

[TH17] T Tieleman and G Hinton. "Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning." *Technical Report.*, 2017.

[TIW17] Jakub M Tomczak, Maximilian Ilse, and Max Welling. "Deep learning with permutation-invariant operator for multi-instance histopathology classification." *arXiv preprint arXiv:1712.00310*, 2017.

[TL19] Mingxing Tan and Quoc V Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." *arXiv preprint arXiv:1905.11946*, 2019.

[TLB19] David Tellez, Geert Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen van der Laak. "Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology." *Medical image analysis*, **58**:101544, 2019.

[TSS18] Syed Ahmed Taqi, Syed Abdus Sami, Lateef Begum Sami, and Syed Ahmed Zaki. "A review of artifacts in histopathology." *Journal of oral and maxillofacial pathology: JOMFP*, **22**(2):279, 2018.

[UVL16] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. "Instance normalization: The missing ingredient for fast stylization." *arXiv preprint arXiv:1607.08022*, 2016.

[VCC17] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. "Graph attention networks." *arXiv preprint arXiv:1710.10903*, 2017.

[VFB16] Gitte Vanwinckelen, Daan Fierens, Hendrik Blockeel, et al. "Instance-level accuracy versus bag-level accuracy in multi-instance learning." *Data mining and knowledge discovery*, **30**(2):313–341, 2016.

[VFP17] Joost JM Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts. "Computational radiomics system to decode the radiographic phenotype." *Cancer research*, **77**(21):e104–e107, 2017.

[VPS16] Abhishek Vahadane, Tingying Peng, Amit Sethi, Shadi Albarqouni, Lichao Wang, Maximilian Baust, Katja Steiger, Anna Melissa Schlitter, Irene Esposito, and Nassir Navab. "Structure-preserving color normalization and sparse stain separation for histological images." *IEEE transactions on medical imaging*, **35**(8):1962–1971, 2016.

[WHH89] Alex Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J Lang. "Phoneme recognition using time-delay neural networks." *IEEE transactions on acoustics, speech, and signal processing*, **37**(3):328–339, 1989.

[WMR16] Jiazhuo Wang, John D MacKenzie, Rageshree Ramachandran, and Danny Z Chen. "A deep learning approach for semantic segmentation in histology tissue images." In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 176–184. Springer, 2016.

[WSL14] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. "Learning fine-grained image similarity with deep ranking." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1386–1393, 2014.

[WSM20] Ellery Wulczyn, David F Steiner, Melissa Moran, Markus Plass, Robert Reihs, Fraser Tan, Isabelle Flament-Auvigne, Trissia Brown, Peter Regitnig, Po-Hsuan Cameron Chen, et al. "Interpretable Survival Prediction for Colorectal Cancer using Deep Learning." *arXiv preprint arXiv:2011.08965*, 2020.

[WYH15] Jiajun Wu, Yinan Yu, Chang Huang, and Kai Yu. "Deep multiple instance learning for image classification and auto-annotation." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3460–3469, 2015.

[WZ00] Jun Wang and Jean-Daniel Zucker. "Solving multiple-instance problem: A lazy learning approach." 2000.

[WZY19] Shujun Wang, Yaxi Zhu, Lequan Yu, Hao Chen, Huangjing Lin, Xiangbo Wan, Xinjuan Fan, and Pheng-Ann Heng. "RMDL: Recalibrated multi-instance deep learning for whole slide gastric image classification." *Medical image analysis*, **58**:101549, 2019.

[XJW17] Yan Xu, Zhipeng Jia, Liang-Bo Wang, Yuqing Ai, Fang Zhang, Maode Lai, I Eric, and Chao Chang. "Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features." *BMC bioinformatics*, **18**(1):281, 2017.

[XWC15] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. "Empirical evaluation of rectified activations in convolutional network." *arXiv preprint arXiv:1505.00853*, 2015.

[XZE12] Yan Xu, Jianwen Zhang, I Eric, Chao Chang, Maode Lai, and Zhuowen Tu. "Context-constrained multiple instance learning for histopathology image segmentation." In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 623–630. Springer, 2012.

[XZE14] Yan Xu, Jun-Yan Zhu, I Eric, Chao Chang, Maode Lai, and Zhuowen Tu. "Weakly supervised histopathology cancer image segmentation and classification." *Medical image analysis*, **18**(3):591–604, 2014.

[YZB16] Kun-Hsing Yu, Ce Zhang, Gerald J Berry, Russ B Altman, Christopher Ré, Daniel L Rubin, and Michael Snyder. "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features." *Nature communications*, **7**(1):1–10, 2016.

[YZP16] Zhennan Yan, Yiqiang Zhan, Zhigang Peng, Shu Liao, Yoshihisa Shinagawa, Shaoting Zhang, Dimitris N Metaxas, and Xiang Sean Zhou. "Multi-instance deep learning: Discover discriminative local anatomies for bodypart recognition." *IEEE transactions on medical imaging*, **35**(5):1332–1343, 2016.

[ZCH19] Niyun Zhou, De Cai, Xiao Han, and Jianhua Yao. “Enhanced Cycle-Consistent Generative Adversarial Network for Color Normalization of H&E Stained Images.” In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 694–702. Springer, 2019.

[ZG02] Qi Zhang and Sally A Goldman. “EM-DD: An improved multiple-instance learning technique.” In *Advances in neural information processing systems*, pp. 1073–1080, 2002.

[ZLV17] Wentao Zhu, Qi Lou, Yeeleng Scott Vang, and Xiaohui Xie. “Deep multi-instance networks with sparse label assignment for whole mammogram classification.” In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 603–611. Springer, 2017.

[ZML07] Jianguo Zhang, Marcin Marszałek, Svetlana Lazebnik, and Cordelia Schmid. “Local features and kernels for classification of texture and object categories: A comprehensive study.” *International journal of computer vision*, **73**(2):213–238, 2007.

[ZPV06] Cha Zhang, John C Platt, and Paul A Viola. “Multiple instance boosting for object detection.” In *Advances in neural information processing systems*, pp. 1417–1424, 2006.

[ZSL09] Zhi-Hua Zhou, Yu-Yin Sun, and Yu-Feng Li. “Multi-instance learning by treating instances as non-iid samples.” In *Proceedings of the 26th annual international conference on machine learning*, pp. 1249–1256, 2009.

[ZWF18] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. “Adversarial complementary learning for weakly supervised object localization.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1325–1334, 2018.

[ZYZ17] Xinliang Zhu, Jiawen Yao, Feiyun Zhu, and Junzhou Huang. “Wsisa: Making survival prediction from whole slide histopathological images.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7234–7242, 2017.

[ZZ04] Min-Ling Zhang and Zhi-Hua Zhou. “Improve multi-instance neural networks through feature selection.” *Neural processing letters*, **19**(1):1–10, 2004.

[ZZ07] Zhi-Li Zhang and Min-Ling Zhang. “Multi-instance multi-label learning with application to scene classification.” In *Advances in neural information processing systems*, pp. 1609–1616, 2007.