COMPUTER AIDED ANALYSIS OF PROSTATE HISTOPATHOLOGY IMAGES

by

JIAN REN

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Electrical and Computer Engineering

Written under the direction of

David J. Foran

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

OCTOBER, 2019

ABSTRACT OF THE DISSERTATION

COMPUTER AIDED ANALYSIS OF PROSTATE HISTOPATHOLOGY IMAGES

By JIAN REN

Dissertation Director: David J. Foran

Prostate cancer is the most common non-skin related cancer affecting 1 in 7 men in the United States. Treatment of patients with prostate cancer remains a difficult decision-making process that requires physicians to balance clinical benefits, life expectancy, morbidities, and potential side effects. Gleason scores have been shown to serve as the best predictors of prostate cancer outcomes. In spite of progress made in trying to standardize the grading process, there still remains approximately a 30% grading discrepancy between the score rendered by general pathologists and those provided by experts while reviewing needle biopsies for Gleason pattern 3 and 4, which accounts for more than 70% of daily prostate tissue slides at most institutions. Therefore, we present computational imaging methods for prostate gland analysis which we will utilize to develop an automated reliable computer-aided Gleason grading system. The inspiration for the project starts from the fact that prostate adenocarcinoma is diagnosed by recognizing certain histology fields clinically. Recently, the Gleason grading criteria used to perform Gleason grading was updated to allow more accurate stratification and higher prognostic discrimination as compared to the traditional grading system.

In this thesis work, we have gone beyond Gleason score analysis by introducing survival model assessment to predict patient outcomes. Using whole-slide images (WSIs) generated from biopsy tissues from radical prostatectomy surgical specimens, we utilize deep learning approaches to discover the most promising computational image biomarkers. The proposed method differs from existing survival analysis studies that use individual patches or manually designed protocols to select a set of patches. In contrast to those approaches, we develop an end-to-end methodology to learn from patches that are analyzed sequentially while preserving their inter-spatial relationships within the WSIs. We build the automatically cropped patches from a WSI as a sequence and use the recurrent neural network to generate a salient representative computational biomarker for the WSI.

Automatic and accurate Gleason grading of histopathology tissue slides is crucial for reliable prostate cancer diagnosis, treatment, and prognosis. Usually, histopathology tissue slides from different institutions show heterogeneous appearances because of variation in tissue preparation and staining procedures, thus the predictable model learned from one domain may not be applicable to a new domain, directly. Here we propose to adopt unsupervised domain adaptation to transfer the discriminative knowledge obtained from the source domain to the target domain without requiring labeling of images at the target domain. The adaptation is achieved through adversarial training to find an invariant feature space along with the proposed Siamese architecture on the target domain to add the regularization that is appropriate for the whole-slide images. We validate the method on two prostate cancer datasets and obtain significant classification improvement of Gleason score as compared with the baseline models.

Finally, we explore the possibility of utilizing cluster computing infrastructure to speed up the analysis. The nuclei detection algorithm that was previously reported extremely reliable in terms of accuracy, but suffered from the fact that performance took an inordinate amount of time to run on a single machine. We have addressed this challenge and present here a parallel nuclei detection algorithm that has been implemented on CometCloud.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my advisor, Dr. David J. Foran, for his guidance and support for the past five years. I am genuinely grateful for his advice, encouragement, patience, and cheerfulness.

I would like to thank Dr. Manish Parashar, Dr. Ilker Hacihaliloglu, and Dr. Eric A. Singer, for serving as my Ph.D. committee members and spending their precious time in reviewing my thesis.

I want to thank all the collaborators for their help and guidance. They include Dr. Zhe Lin, Dr. Xiaohui Shen, Dr. Jianchao Yang, Dr. Ning Xu, Dr. Chen Fang, Dr. Radomir Mech, Dr. Michael Gatza, and Dr. Evita Sadimin. I would also like to thank all my friends and colleagues at Rutgers for their generous support and help during my Ph.D. study.

My deepest gratitude goes to my parents Rong Ren and Hongling Yu, my wife Yue Yang, and my daughter Lila, for their understanding and love.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Prostate cancer is the second common cancer among men in the United States, and the second leading cause of cancer death in American men, according to the latest statistics from the American Cancer Society reported in 2017 [1]. Clinical factors including the prostate-specific antigen (PSA) blood test value, patient's age, tumor stage and prostate biopsy grading *et al* are important prognostic features for prostate cancer early detection and diagnosis [1, 2, 3]. After the biopsy of prostate tissue, the prostate cancers are graded according to the Gleason system that assigns a Gleason score based on cancerous cells fall into 5 distinct patterns as they change from normal cells to tumor cells. The cell patterns are graded as a scale of 1 to 5, pattern 3 consists of infiltrative well-formed glands, varying in size and shapes, pattern 4 consists of poorly formed, fused or cribriform glands, pattern 5 consists of solids sheets or single cells with no glandular formation. The Gleason grading system has been shown to be the strongest prognostic factor for men with prostate adenocarcinoma.

The Gleason score is solely based on prostate glandular morphological architectures, which is a sum of primary and secondary Gleason patterns exhibited in the tissue pathology image. The newly established prostate cancer grading system which has been developed by experts in the field, features a five-grade group system (group 1 to 5 as Gleason score ≤ 6, $3 + 4$, $4 + 3$, 8 and $9 - 10$ respectively). Generally speaking, prostate cancers with lower Gleason scores (2-4) tend to be less aggressive while prostate cancers with higher Gleason scores (7-10) tend to be more aggressive. The prostate Gleason score remains one of the best predictors for determining risks of prostate cancer progression and predicting patient outcome [4, 5, 6, 7, 8, 9]. Patients with Gleason score

of 7 are divided into two prognostic groups, group II for those with primary pattern 3 + secondary pattern 4, and group III for those with primary pattern 4 + secondary pattern 3 [10]. The most conflicting group of the prognostic difference is Gleason score 7 on a biopsy depending on whether the primary Gleason pattern is 3 or 4. Numerous studies have demonstrated that in radical prostatectomy specimens, Gleason score 4 + 3 has a worse prognosis than 3 + 4 [11, 12, 13, 14, 15, 16, 17, 18]. Since there is a significant difference between pattern 4 + 3 and pattern 3 + 4, it is very important to be able to separate pattern 3 and pattern 4 accurately. Unfortunately, since it is difficult at times to objectively assign these patterns, a substantial interobserver variability exists, especially among general pathologists who do not specialize in urologic pathology [19].

Furthermore, although many prediction tools [20, 21, 22, 23, 24, 25, 26] use whole-slide images (WSIs) of biopsy tissue from radical prostatectomy surgical specimens to assess prostate cancer progression risk and predict the likely outcomes resulting from treatments, no reliable tools yet exist that simultaneously consider clinical factors and tissue WSIs to stratify prostate patients into subgroups with different risks of progression. Therefore, it is important to predict the prostate cancer progression using computational image biomarkers discovered from WSIs. The WSIs are scanned from the biopsy tissues of the radical prostatectomy surgical specimens. Considering the high computational cost on the giga-pixel tissue WSIs, existing WSIs classification and survival analysis approaches are focused on effectively utilizing the cropped patches from region of interests (ROIs) [27, 28, 29, 30, 31]. However, the process of labeling ROIs is labor-intensive and requires expert pathologists to review the ROIs under different magnifications. In addition, the ROIs represent only partial information within the WSIs, especially for the prostate images, where the Gleason grading is a sum of the primary and secondary Gleason patterns for the entire tissue sample. Using the WSI direcitly could preserve more information. Survival analysis is a very useful tool in predicting patient outcome and provides invaluable information regarding intervention. There are three primary goals of survival analysis: estimating and interpreting survival and/or hazard functions from patients' survival data; comparing survival and hazard functions; and assessing the relationship of explanatory variables to survival time.

Given that fact that histopathology WSIs obtained from different institutions usually present distinct glandular region distributions due to differences in appearance that may be caused by using different microscope scanners and staining procedures, therefore such differences may render the supervised classification model used for predicting the Gleason grade for one annotated dataset (source domain) ineffective on another prostate dataset (target domain). A widely used approach to address the challenge is to label new images on the target domain and fine-tune the model trained on source domain [32]. Instead, methods that can learn from existing datasets and adapt to new target domains, without the need for additional labeling, are highly desirable. With the development of unsupervised domain adaptation [33, 34], it is possible to classify the newly given prostate datasets into low and high Gleason grade through unsupervised learning, which could same lots of time and money for labeling WSIs.

Considering the highly computational cost of many computer aiaided algorithms, such as the robust nuclei segmentation algorithm has been reported in [35], which includes two main sequential steps, seed detection and contour generation, it's not very efficient to run nuclei segmentation algorithm directly on the whole image which may contains hundreds and thousands of nuclei. To accelerate the process, there have been many applications using cloud computing on medical image analysis, but most of them were focused on data parallelization instead of the algorithm parallelization [36, 37, 38, 39]. Thus we address the challenge of working with specimens which have not been enhanced with specialized staining methods and can be used across a broader number of application areas.

1.2 Outline

In Chapter 2, we present two computer aided analysis approaches for prostate gland segmentation using pattern 3 and 4 Hematoxylin and Eosin (H&E) stained pathology images. The first one is a region-based nuclei and gland grouping approach, it utilizes the structure information to help segment each gland [25]. The second method that we propose is a two-phase gland classification method. The classification of each gland is based on the accurate segmentation of glandular regions on Hematoxylin and Eosin

(H&E) stained images [26].

In Chapter 3, we conduct Gleason score-guided prostate cancer progression analysis using deep learning approaches on WSIs of biopsy tissues and survival models to develop a higher discriminative and predictive way in patients' outcomes. The prostate cancer patients' disease-free time (months) since their initial treatment are applied as the time-to-recurrence for progression analysis using survival models. we adopt a recurrent neural network (RNN) model [40], namely the long short-term memory (LSTM) network [41] to learn the fine-grained discriminative information among patches (e.g. Gleason pattern 3 and pattern 4) and the global representations of the WSI. Unlike the traditional RNN that has vanishing and exploding gradients problem [40], LSTM incorporates memory cells with several gates to obtain long-range dependencies by enabling the network to learn at what time to forget previous hidden states as well as update hidden states with new information. For one WSI, we systematically forward the cropped patches into CNN and get the activations from the second to the last layer. On top of that, the LSTM network maps the sequence of the activations into one feature vector that encodes the global representative information of the WSI.

Futhermore, in chapter 4, we build a unified system using public available whole-slide images and genomic data of histopathology specimens through deep neural networks to identify a set of computational biomarkers. Using a survival model, experimental results on the public prostate dataset showed that the computational biomarkers extracted by our approach had hazard ratio as 5.73 and C-index as 0.74, which were higher than standard clinical prognostic factors and other engineered image texture features. Collectively, the results of this study highlight the important role of neural network analysis of prostate cancer and the potential of such approaches in other precision medicine applications [42, 43].

In Chapter 5, we propose a novel Factorized Adversarial Networks to tackle the unsupervised domain adaptation in an effective way [44]. Furthermore, we adopt the domain adaptation for unsupervised prostate histopathology WSIs classification in Chapter 7. We apply adversarial training to minimize the distribution discrepancy at the feature space between the domains, with the loss function adopted from the Generative

Adversarial Network (GAN) [45]. Furthermore, we developed a Siamese architecture for the target network to serve as a regularization of patches within the WSIs. The proposed method is validated on public prostate datasets and a newly collected local dataset. The experimental results show the approach significantly improves the classification accuracy of Gleason score as compared with the baseline model. To the best of our knowledge, this is the first study of domain adaptation for unsupervised prostate histopathology WSIs classification [46, 47].

In Chapter 7, we propose a new approach to parallelize the nuclei detection algorithm by utilizing CometCloud to speed up the whole process to make the nuclei segmentation running in real-time a possibility.

Chapter 2

Patch-based Gland Segmentation and Classification

2.1 Introduction

With the rapid development and adoption of whole-slide microscopic imaging and the corresponding advances being made in terms of available computing power, the potential for developing a reliable, automated computer-aided diagnosis (CAD) system capable of performing objective, reproducible Gleason scoring while avoiding intra- and inter-observer variability is now technically feasible. The newly established prostate cancer grading system which has been developed by experts in the field, features a five-grade group system. This methodology offers more accurate grade stratification than traditional systems and provides the highest prognostic discrimination for all cohorts on both univariate and multivariate analysis [4].

There have been many studies on computer-aided Gleason grading, however most of them are not focused analyzing intact glandular regions. In general there are four approaches on prostate Gleason pattern grading including color-statistical based [48], texture-based [49, 50], structure-based [51], and tissue-component-based [52, 53]. To achieve significant improvements in discriminating between Gleason score 3 and 4, it is essential to first perform accurate segmentation of individual glandular regions.

In the following, we describe two computational imaging decision support frameworks which are investigated as a deployable tool to allow accurate discrimination among even the most challenging Gleason patterns 3 and 4 in prostate cancer diagnoses. The first method is a region-based nuclei segmentation to get individual gland without using lumen as prior information. The second one is a two-phase gland classification method. The classification of each gland is based on the accurate segmentation of glandular regions on Hematoxylin and Eosin (H&E) stained images.

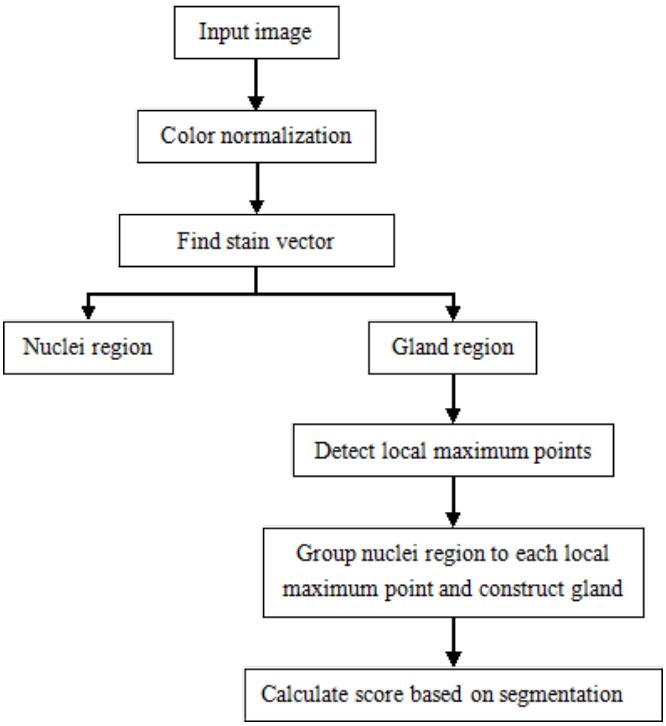


Figure 2.1: Flow chart of region-based nuclei segmentation.

2.2 Region-based Nuclei Segmentation

The region-based nuclei and gland grouping approach segments individual gland by grouping its surrounding nuclei without using lumen as prior information. The method includes three main steps. First, image pre-processing is implemented for removal of staining variations from different images. Second, all the nuclei and global gland regions are identified respectively. Third, each individual gland is constructed from the distance map of gland region with grouping of its adjacent surrounding nuclei.

The several steps of our approach are summarized in Figure 2.1. Using a well-defined H&E stained image as a reference image, all the images are normalized for removal staining variations. From their staining vectors, nuclei and glands regions are identified by color deconvolution. Based on an assumption that only one whole lumen locates in one single gland no matter grade 3 individual gland structure or grade 4 gland infusion, local maximum points in distance map of gland region are applied to identify the number of glands in each gland region. Because many glands do not have lumen and lumen usually located in the center of glands, surrounding adjacent nuclei

on each gland region are classified to different local maximum points by Delaunary triangulation grouping approach.

2.2.1 Color Normalization for removal stain variations

Because of stain variations within those H&E stained images, in which each comes from different patient, color normalization is applied to have image quality control as its pre-processing step. Figure 2.2 shows four examples of H&E stained prostate pathology slides. They have different staining appearances for nuclei, cytoplasm, stroma, and lumen, even come from the same institute. We use color map normalization method, which is described in[54]. The reason to choose this normalization approach is that it uses unique color in the image instead of color frequency of all the pixels. So we use a well-defined H&E stained image as a reference image, and all the images are normalized by the color map of the reference image for removal staining variations.

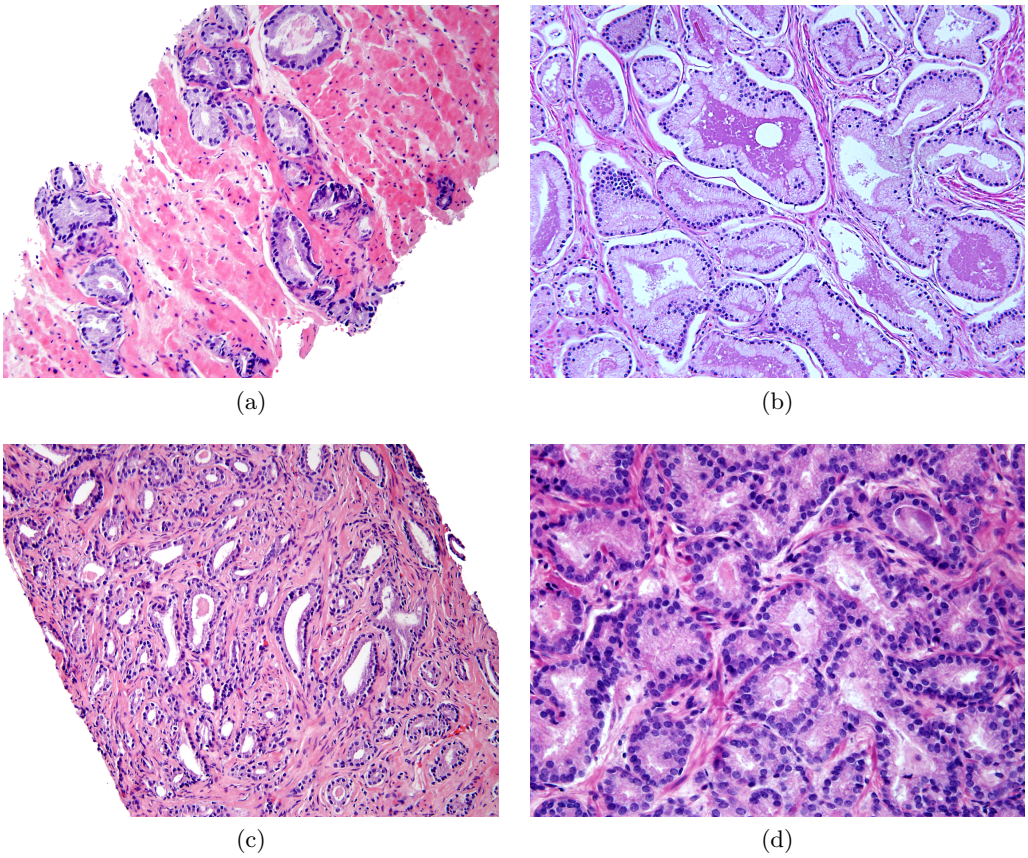
2.2.2 Identification nuclei region and gland region

After color normalization, color deconvolution [55] is applied to extract nuclear region mask and glandular region mask.

Because stain vectors are automatically acquired from color deconvolution, shown as two examples in Figure 2.3 (a). Given a different threshold to the stain vector image, we can get nuclear region mask and glandular region mask. For the glandular region mask, a small area threshold is used to remove noise, as shown in Figure 2.3 (b). And the nuclear region mask is shown in Figure 2.3 (c).

2.2.3 Gland construction

Since we get the mask of nuclear region and glandular region, we can group each nuclei region to each gland, and therefore construct each gland from gland regions. Assume we have a gland region G, and use this gland region multiplies nuclei region to get nuclei regions on the gland, suppose N_k, (k=1...m), where N_k denotes the k^{th} nuclei region on the gland and m denotes total number of nuclei regions on the gland. Figure 2.4



(a) (b)

(c) (d)

Figure 2.2: Examples of four H&E stained images, but with quite different staining appearance for nuclei, cytoplasm, stroma and lumen. (a) and (b) shows a trend of gland infusion, many glands have touched other glands, while glands in (c) and (d) have merged together and it's more difficult than (a) and (b) to separate each gland.

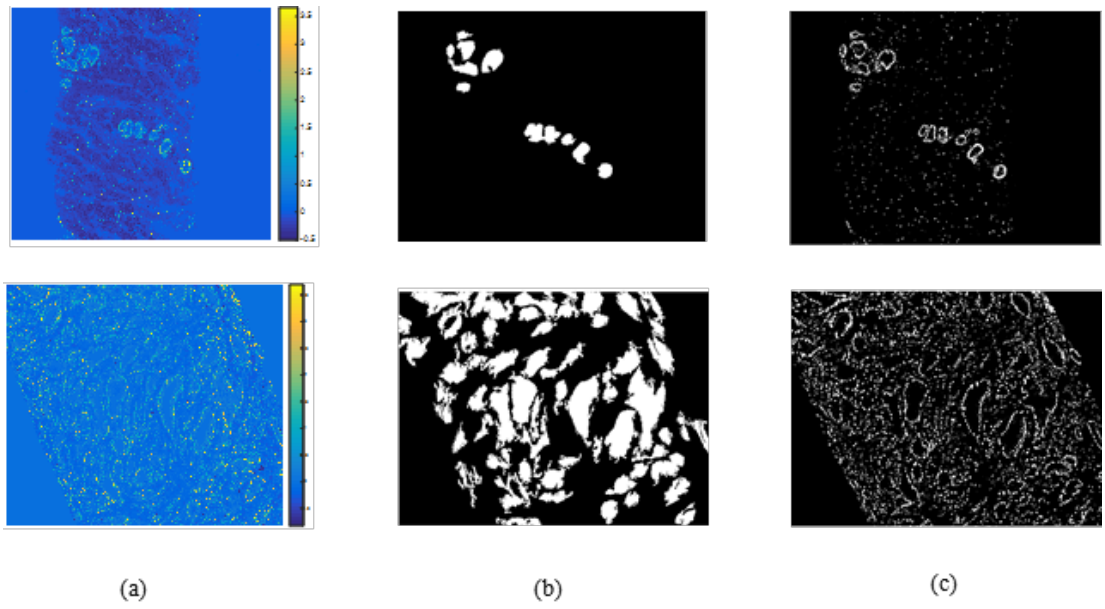


Figure 2.3: (a) Stain vector contains gland and nuclei information; (b) Glandular region mask; (c) Nuclear region mask.

(a) is original image and Figure 2.4 (b) is the glandular region mask of original image and there are three gland regions in the image. The gland construction includes three steps.

Firstly, we transform the gland region to a distance image, then localize its local maximal points on it. The distance map is shown in Figure 2.4 (c), and local maximum points are shown in Figure 2.4 (d). Based on the assumption that only one whole lumen locates in one single gland no matter grade 3 individual gland structure or grade 4 gland infusion, the local maximum points in distance map of gland region are applied to identify the number of glands in each gland region. So we have several glands G_i, (i=1...n), here n denotes number of local maximum points in each gland region and G_i denotes the i^{th} gland region. Figure 2.4 (d) shows the local maximum points on the distance transform map.

Secondly, for the gland region with just one local maximum point, we consider it as a single gland and use nuclei regions' centroids and the local maximum point to construct Delaunay triangulation and therefore find border of the gland; while for others, the Algorithm 1 is applied to each gland region. Since each gland G_i has its

nuclei regions, we can construct part of the complete gland by using local maximum point and nuclei regions, as shown in Figure 2.4 (e).

Algorithm 1: Gland construction of region-based nuclei segmentation

```
while While there is nuclei region N_k unlabeled do
    if The nuclei region N_k only directly connected with G_i then
        N_k is classified as G_i's nuclei
    if The nuclei N_k directly connected with multiple local maximum points, such
    as G_p ... G_q (1 ≤ p < q ≤ n) then
        N_k is classified to the gland having the largest nuclei density
    if The nuclei densities of G_p ... G_q are same then
        Choose the gland with closest distance
```

Thirdly, for those non-directly connected nuclei regions, searching their adjacent classified glands, they are assigned to the gland with smallest distance.

The glands have less than three nuclei region will be discarded because it's not enough to construct a triangulation. After all the nuclei regions are grouped, we reconstruct each gland using Delaunay triangulation, the result is shown in Figure 2.4 (f), and each black arrow denotes a segmented gland.

2.3 Two-Phase Gland Classification

The first part of two-phase gland classification is delineating each image by using the segmentation network to generate an image mask. We use semantic pixel-wise classification to get the binary mask of input RGB image. The segmentation networks includes encoding the image and then decoding it. In the second phase, the features abstracted from each segmented gland are subsequently used as the inputs for a random forest and a score between 3 and 4 is given for each gland. Experimental results show that the two-phase classification approach achieves improved prostate glandular segmentation and classification results on H&E stained images compared to state-of-the-art.

2.3.1 Prostate image segmentation

The segmentation network that we have developed is based on a convolution neural network (CNN) which can be trained end-to-end with stochastic gradient descent to

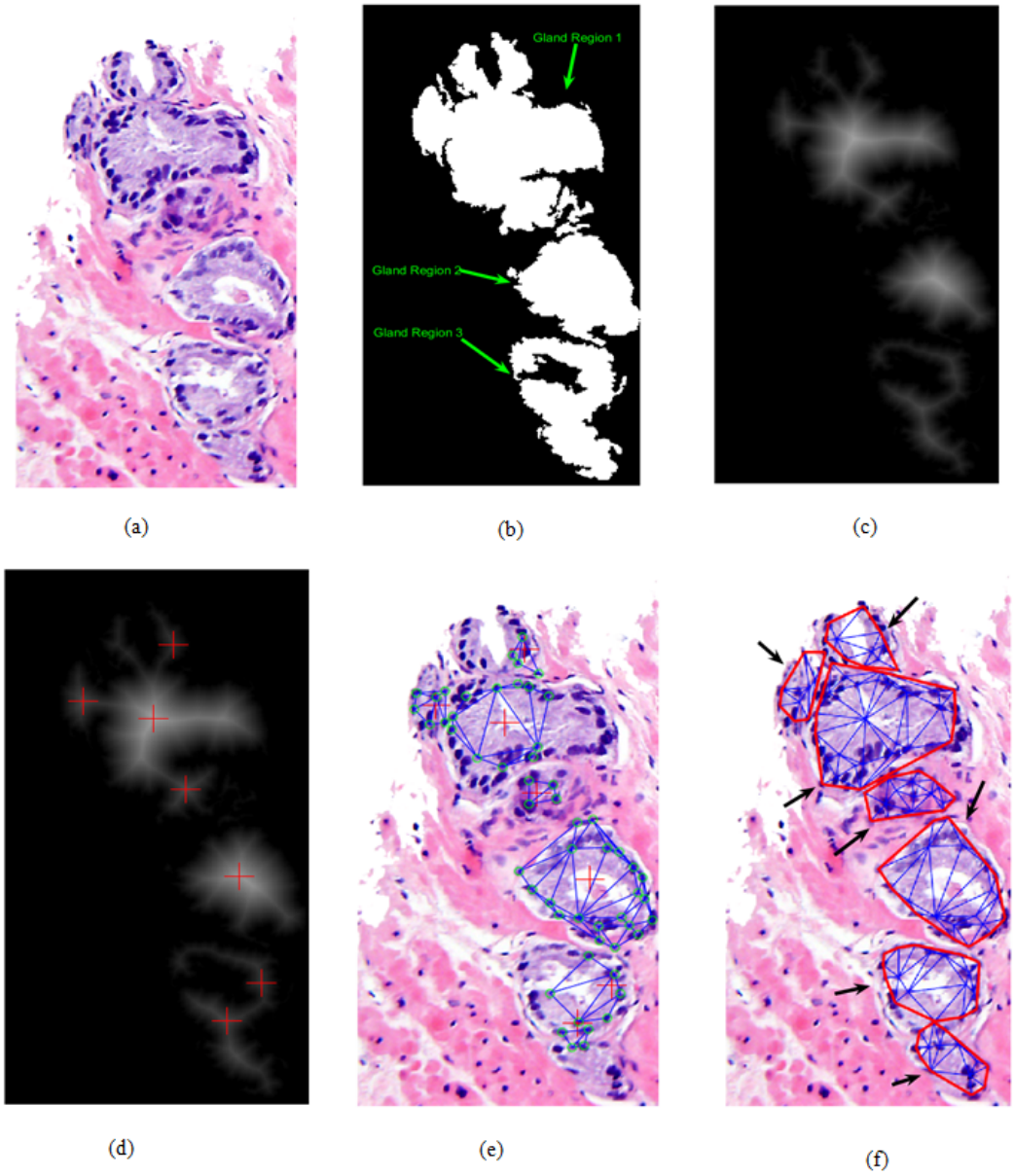


Figure 2.4: (a) Original Image; (b) Gland region mask, in which there are three gland regions in the images and are labeled by green arrows; (c) Distance transform of glandular region mask; (d) Local maximum points of distance transform image; (e) Grouping the nuclei that connect with local maximum points directly; (f) Contours of final (in red) image of gland segmentation, each black arrow indicates one gland.

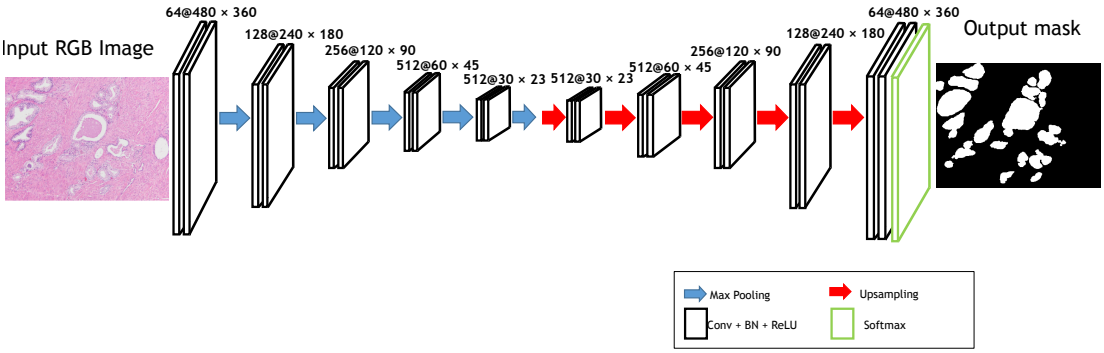


Figure 2.5: The architecture of the semantic segmentation network.

give the semantic pixel-wise segmentation of the original input RGB images. As shown in Figure 2.5, CNN consists of encoding and decoding module but does not contain a fully connected layer. Both the encoding portion of the network and the decoding component contain 10 convolutional layers. The encoding part includes the typical convolutional network and the convolutional layers are composed of kernel size 3×3 and padding size 1 and are followed by a rectified linear unit (ReLU) $max(0, x)$, batch normalization (BN) layer[56] and 2×2 max pooling layer with stride 2. The max pooling layer is replaced by the upsampling layer[57] in the decoding component of the network. The upsampling layer uses the location from the max pooling layer to reverse operation of max pooling with stride 2. The final layer is the soft-max classifier for the binary classification with the cross-entropy loss function as the objective function to train the network.

In order to retain the boundary information during the test phase, each image is mirrored by the four boundaries as shown in Figure 2.6. In this manner, the center of each output image can be utilized to form the seamless segmentation mask and the mask has the same size as the test image. Morphological operations are used as a post-processing step to remove artifacts.

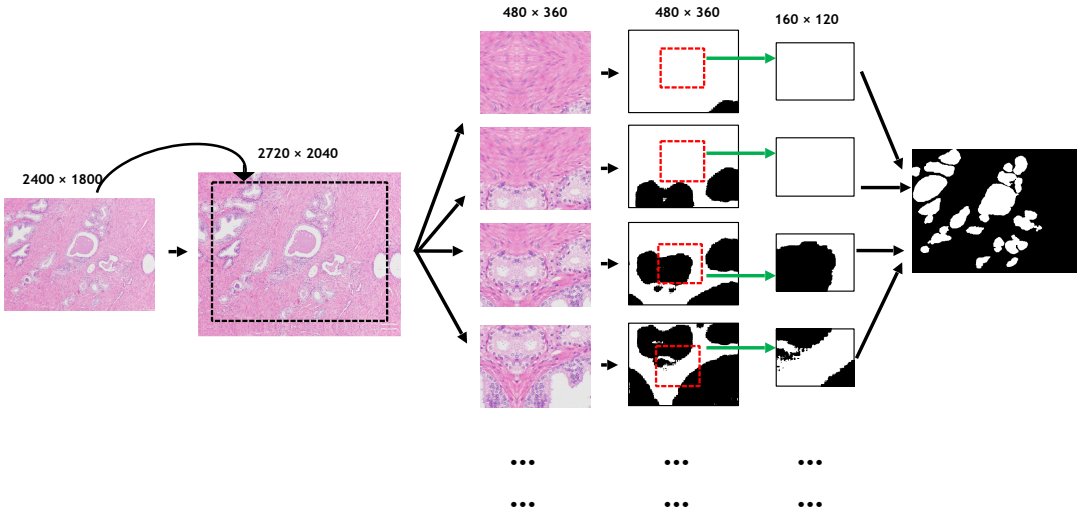


Figure 2.6: Each test image is mirrored by four boundary sub-images in order to retain the boundary information. And each test image is cropped into several sub-images. Only the center of each predicted sub-image mask is kept to form the preliminary mask.

2.3.2 Gland Grading based on segmentation

Superpixel Segmentation

For the Gleason pattern 3 glands, the lumen is typically surrounded by nuclei. While glands begin to merge or fuse together in the Gleason pattern 4 glands, the lumen may not be surrounded by nuclei and their spatial co-localization could be an arbitrary pattern. Therefore we take advantage of the spatial structure pattern to differentiate Gleason pattern 3 and 4. Using superpixel segmentation method[58], the segmented glands from above step is then segmented into two sub-images: (1) the outer boundary image and (2) the inner center image. The segmented region S_i is classified to the boundary image if they are adjacent to the background. Suppose the number of segmented regions in the boundary image is m. Then the center of the original image is extracted from the distance map. If there are m nearest superpixel regions adjacent to the center, those m nearest regions form the center image. If the left superpixel regions are less than m, all of them form the center image. An illustration of segmentation of outer boundary image and inner center image is shown in Figure 2.7.

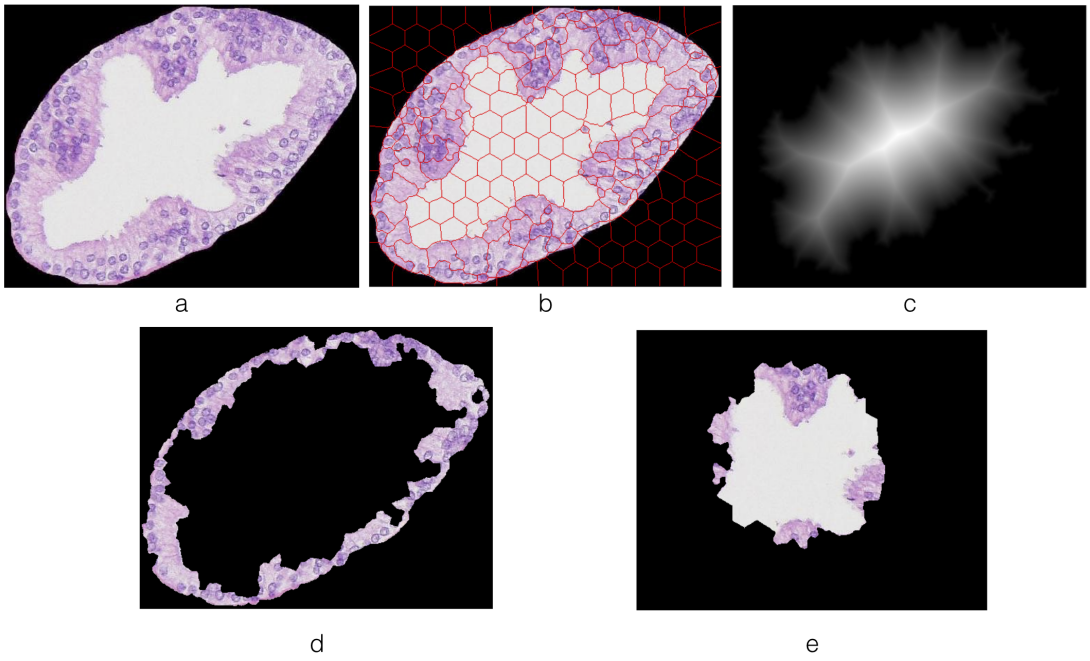


Figure 2.7: (a) original image; (b) superpixel segmentation on the original image; (c) distance map of the original image; (d) image contains boundary information; (e) image contains center information.

Feature Extraction

Texture, shape and color features are extracted from the boundary images and the center images to train the random forest classifier. The texture features are calculated by using Bag-Of-Word on SIFT features. SIFT texture features are extracted from training images and clustered by K-means algorithm. Using Bag-of-Word paradigm, each image has k-bins of spatial histogram of K-means cluster centers as its texture features. Here we use K equals 300 in our experiments after different K value testing. The shape descriptor in each image is represented by HOG features. And we use mean, standard deviation and the 5-bin histograms of intensities for each R, G, B channel to represent the color feature. All the texture, shape and color features are consolidated together. Suppose the set of features from the boundary image is represented by f_i^b and the set of features from the center image is represented by f_i^c. To enhance the difference between the boundary image and center image, we use $f_i = \frac{w \times f_i^b}{(1-w) \times f_i^c}$ to represent the features of the original gland image. w is is a weight parameter, varying from 0.1 to 0.9.

Random Forest Regression

The grading of each gland between 3 and 4 is based on random forest regression. A random forest is an ensemble of a number of decision trees, with each tree trained using a randomly selected training sets. The output of a decision tree is produced by branching an input left or right down the tree recursively until meet any leaf node. The decision forest combines the predictions from individual tree using an ensemble model and gives the regression output by averaging. The output score of the test image should be in the range of 3 to 4.

2.4 Experiment Results

In this study, all the prostate images were from Pathology Department at Johns Hopkins Medical Institutes. The images were stained by H&E. Our experiments consist of 22 prostate images from 22 difference patients. The images are under 20× magnification with a size of 2400×1800.

Here we analyze the time complexity for the region-based nuclei segmentation. For the glands which only have one local maximum points, the time complexity is $O(X_1)$, where X_1 is the number of glands with one local maximum points; while for other glands, the time complexity is $O(X_2(N^2 + M^2))$, where X_2 is the number of glands with multi local maximum points, N is the number of nuclei regions on the glands and M is the number of local maximum points. For each image, using a computer with the Intel Xeon processor and 16.0 GB RAM, the approach is implemented in MATLAB and the average running time is less than one minute.

For the two-phase gland classification, we use 5-fold cross-validation to randomly select 17 images as training images and others as testing images for the segmentation network. 25 images are cropped from each image and the size of the cropped image is 480×360. Each cropped image is horizontal flip and vertical flip, so 1275 images are used to train the image segmentation network. Precision (P), recall (R) and F_1 score are used to measure the segmentation quantitatively. P is denoted as the intersection between the segmentation results and the manually annotation results divided by the

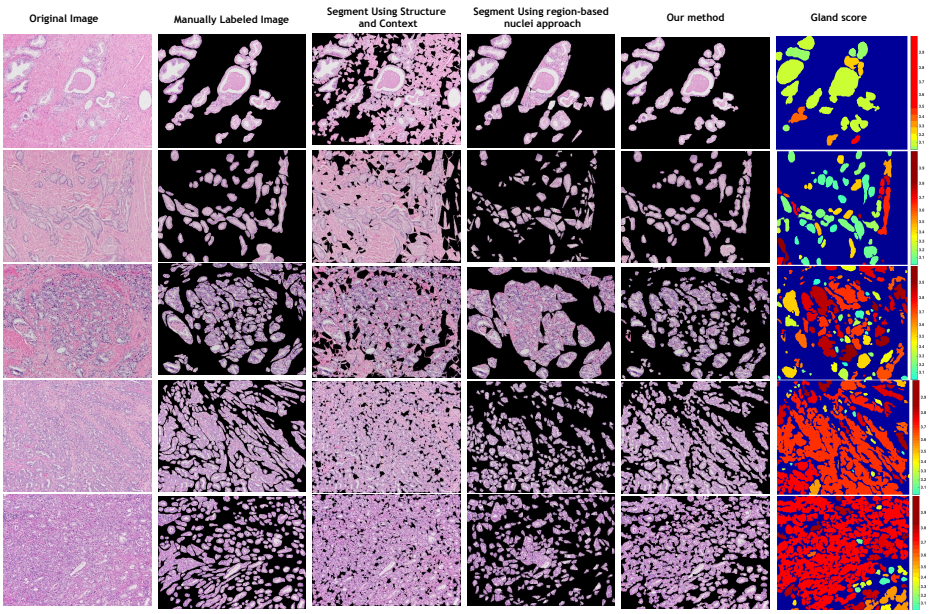


Figure 2.8: Results are shown for different methods. A score is given for each gland after segmentation.

segmentation results while R is divided by the manually annotation results. So we can have $F_1 = \frac{2\times P\times R}{P+R}$. We achieve F_1 score as 0.8460 for the segmentation network. Table 2.1 shows the segmentation performance comparison for different methods. The segmentation network is implemented by using Caffe [59] on NVIDIA Quadro K5200 GPU with cuDNN acceleration.

After the each gland is segmented, we use the 634 labeled glands to train the random forest classifier. All these glands are obtained from the 22 H&E stained images. Each gland image is resized as 360×360. The weight parameter w for the feature exaction equals to 0.7 for the best classification accuracy and the number of trees in the random forest is 160 for a stable regression score. We use 10-cross validation to perform the training. The sensitivity, specificity and accuracy for the classification are 0.70±0.15, 0.89±0.04 and 0.83±0.03 respectively. Figure 2.8 shows the segmentation results for different methods and the scores given for each gland after the segmentation.

Table 2.1: Segmentation Performance Comparison for Different Methods

	Precision	Recall	F_1 Score
Structure and Context[60]	0.4748	0.9530	0.6224
Region-based Nuclei Approach	0.8103	0.6703	0.7175
CNN without post-processing	0.8823	0.8235	0.8453
CNN with post-processing	**0.8921**	0.8123	**0.8460**

2.5 Conclusion

In this chapter, we present two methods for quantitatively analyzing histopathology prostate cancer images representative of Gleason pattern 3 and 4. The computer-aided analysis framework that we developed for performing prostate Gleason grading achieves a better segmentation result compared to the state-of-the-art approaches. Meanwhile it provides a quick reliable means for grading glandular regions especially those types more often found in Gleason pattern 4. Based on these results, the methods described may lead to a more reliable approach to assist pathologists in performing stratification of prostate cancer patients and improves therapy planning.

Chapter 3

Prostate Cancer Progression Analysis using Computational Pathology Whole-Slide Images

3.1 Introduction

Survival analysis is a means for predicting patient outcomes, by providing invaluable information for selecting treatment. Predicting prostate cancer survival outcomes is a significant challenge. Following radical prostatectomy, men must be closely monitored for evidence of recurrence. This is typically done via prostate-specific antigen (PSA) blood tests. A detectable or rising PSA after surgery is evidence of biochemical recurrence. The measure of time from surgery to biochemical recurrence is biochemical recurrence-free survival (bRFS). Multiple studies have examined predictors of bRFS using quantitative histopathology features with some survival models [61, 62, 63, 64]. Although numerous prediction tools [20, 21, 22, 23, 24, 25, 26] utilized whole-slide images (WSIs) to assess prostate cancer recurrence and predicted the likely outcomes resulting from treatments, several of these studies simultaneously considered clinical factors (primary and secondary Gleason patterns, PSA value, age, tumor stage) and tissue WSIs to correlate with recurrence under different survival models.

The Gleason scoring system for prostate cancer remains one of the best predictors for prostate cancer progression and recurrence [5, 6, 7, 8], despite significant inter-observer reproducibility among pathologists [19, 65, 66]. A more recently adapted grading system stratifies patients into 5 prognostic grade groups [4] based on their Gleason patterns: grade group 1 (Gleason $\leq$3+3=6), grade group 2 (Gleason 3+4=7), grade group 3 (Gleason 4+3=7), grade group 4 (Gleason 4+4=8, 3+5=8 and 5+3=8), and grade group 5 (Gleason 4+5=9, 5+4=9 and 5+5=10). Figure 3.1 shows an example of giga-pixel whole-slide image with different Gleason patterns. The green framed patch

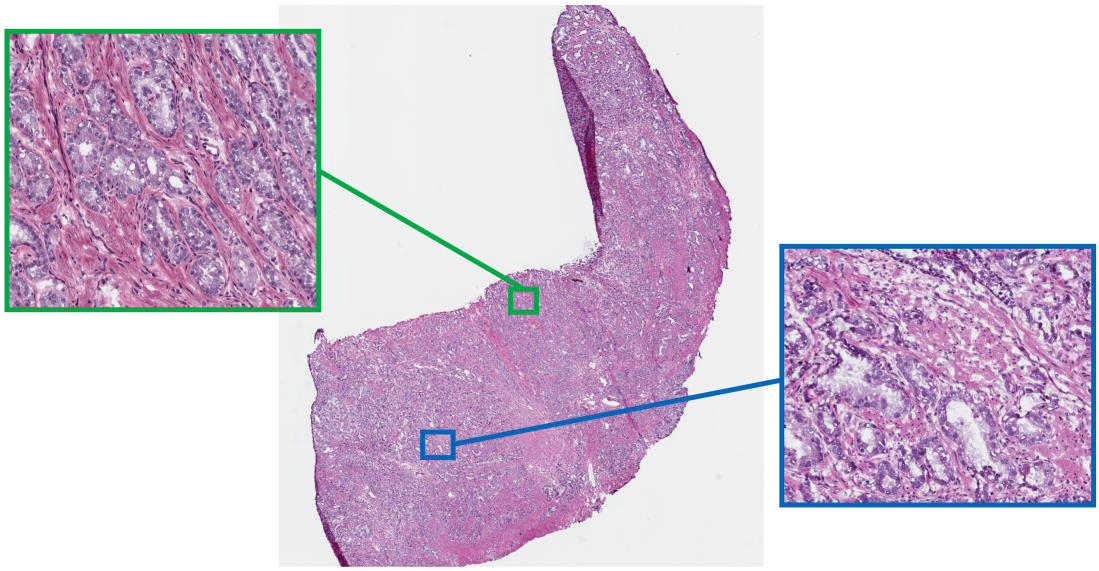


Figure 3.1: Example giga-pixel whole-slide images. The green framed patch is Gleason pattern 3 section and the blue framed patch is Gleason pattern 4.

contains Gleason pattern 3; the blue frames patch contains Gleason pattern 4 and the red frames patch contains Gleason pattern 5. In this study, we conducted experiments on the public prostate cancer dataset using different feature quantification methods and recurrence analysis using the different survival models. Histopathology image features were quantified through texture methods and neural network-based approaches. We focused on the prostate cancer grade groups of 1 to 4. The biochemical bRFS was applied as the time-to-recurrence for prostate cancer progression analysis.

3.2 Materials

In this study, we used the prostate dataset from The Genomic Data Commons (GDC) [67]. The dataset included whole-slide histopathology images from patients and their corresponding clinical reports including the primary and secondary Gleason pattern, patients' prostate-specific antigen (PSA) value, age, and tumor stage. All the image data, annotations of Gleason score, and clinical information were publicly available.

We selected the patients with low-risk (Gleason score 3+3), intermediate-risk (Gleason score 3+4 or 4+3), and high-risk prostate cancer (Gleason score 4+4) because those

Table 3.1: The number of WSIs and their corresponding automatically selected patches under different Gleason scores composing from a sum of Gleason patterns 3+3, 3+4, 4+3 and 4+4 prostate prognostic grading groups.

Gleason Score	3 + 3	3 + 4	4 + 3	4 + 4
# WSIs	43	144	99	49
# patches	1229	4753	2997	1597

patient populations show a reasonable range of prognoses for our analysis. We excluded patients with Gleason grade group 5 patients in this study due to the poor prognosis of their disease [68]. Considering the high computational cost on the giga-pixel tissue WSIs, existing WSIs classification and recurrence analysis approaches were focused on effectively utilizing the cropped patches from region of interests (ROIs) [27, 28, 29, 30, 31]. For image preparation, we adopted the two-step cropping-selecting process. First, original patches were automatically generated within each WSI under 40× objective magnification with a patch size of 4096×4096. Second, the patches with the tissue accounting for at least 20% of the whole area were selected for our experiments. The number of WSIs and cropped patches under different Gleason scores are shown in Table 3.1.

3.3 Methods

Initially, we utilized various quantification methods to extract image features from WSIs. Next, the recurrence analysis was performed on the combination of image features and clinical factors utilizing different survival models, as shown in Figure 3.2. Hazard ratios using different survival models were calculated to indicate correlation between image features (or in context of clinical factors) and recurrence.

3.3.1 Image Feature Quantification

We adopted five approaches for the purpose of feature quantification including unsupervised and supervised methods. The unsupervised texture methods consisted of speeded-up robust features (SURF) [69], histogram of oriented gradients (HOG) [70],

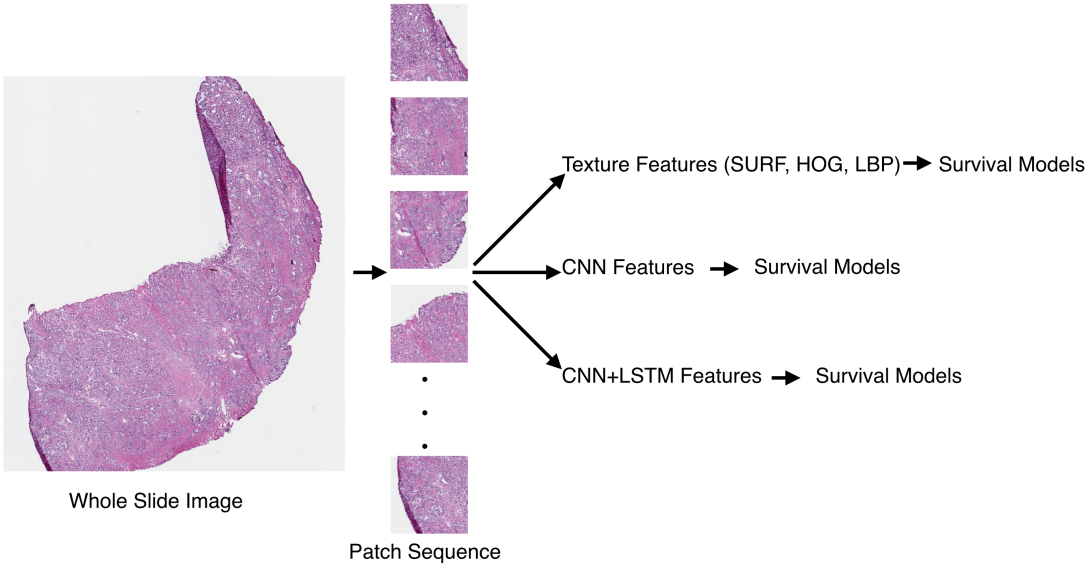


Figure 3.2: Different image features are extracted from WSIs and assessed by various survival models.

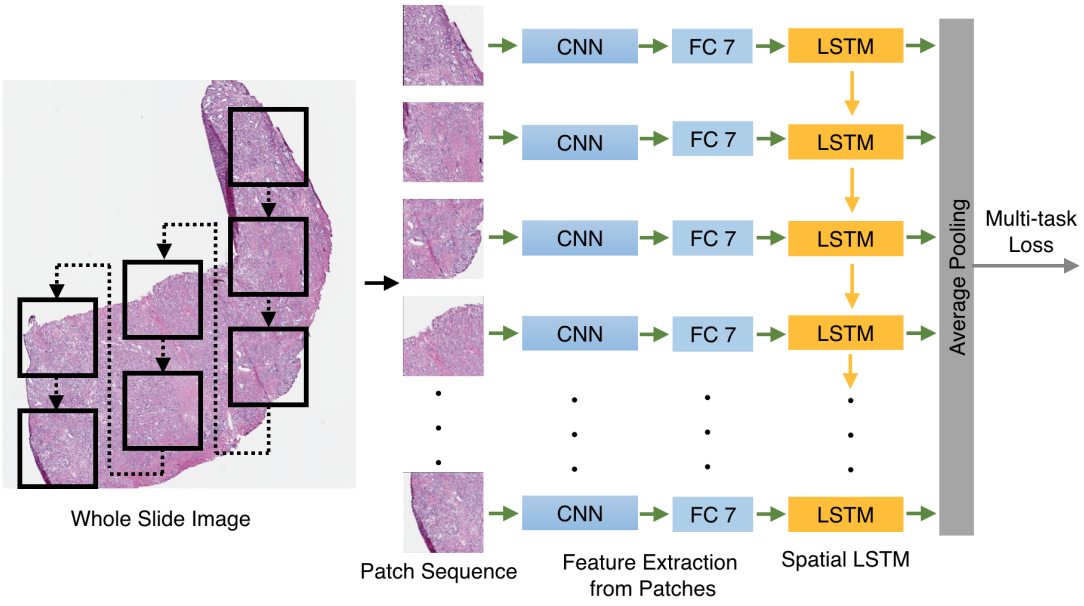


Figure 3.3: The multi-task neural network architecture for computational image features extraction from WSIs. The cropped patches are formed as a sequence by the image coordinates. The LSTM is built on top of the convolutional neural network for the long-term spatial modeling of the activation sequence. An average pooling layer maps the activations into one feature vector.

and local binary pattern (LBP) [71]. The two supervised methods are based on convolutional neural networks. For supervised methods, we randomly selected 20% of the cases as testing set, 10% as validation set and the remainder as training set.

Texture Features

We chose three texture methods for prostate cancer histopathology image analysis. They were rotation, translation, scale-and intensity-invariant which were suitable for descriptions of the texture features within WSIs.

The SURF [69] is partly inspired by the scale-invariant feature transform (SIFT) descriptors. The standard version of SURF is several times faster than SIFT and more robust against different image transformations than SIFT. The image is transformed into coordinates, using the multi-resolution pyramid technique, to copy the original image with Pyramidal Gaussian or Laplacian Pyramid shape to obtain an image with the same size but with reduced bandwidth. The HOG [70] counts occurrences of gradient orientation in a local region of an image. It is similar to that of edge orientation histograms, scale-invariant feature transform descriptors, and shape contexts, but differs in that it is computed on a dense grid of uniformly spaced cells and uses overlapping local contrast normalization for improved accuracy. The LBP [71] is used to model the image local features in texture spectrum units in a multi-resolution gray-scale mode. It is based on recognizing the certain local binary unit patterns for any quantization of the angular space and spatial resolution.

The image features for each patch are generated by using bag-of-word approach [72] from the texture features of different texture methods respectively. By treating image features as words, a bag of words is a sparse vector of occurrence counts (histogram) of a vocabulary of local image features. In the bag-of-word approach, it converts vector-represented texture features to codewords, which also produce a codebook. The image features are mapped to certain codewords through the clustering process and the image is then represented by the histogram of the codewords. Empirically, we use 100 as the number of cluster centers to report the best performance for texture features. In order to select the texture features for WSIs, we apply principal component analysis

Table 3.2: The convolutional neural network applied in our approach. All the convolution layers (Conv) are followed by Rectified Linear Units (ReLU). For the fully connected layers (FC), the FC6 and FC7 are followed by the ReLU and dropout layer with the dropout ratio as 0.5; FC8 and FC9 are both at the top of FC7.

Layer	Filter size, stride	Output W × H × N
Input	-	256 × 256 × 3
Conv	11 × 11, 4	55 × 55 × 96
Max-pooling	3 × 3, 2	27 × 27 × 96
Conv	5 × 5, 1	27 × 27 × 256
Max-pooling	3 × 3, 2	13 × 13 × 256
Conv	3 × 3, 1	13 × 13 × 384
Conv	3 × 3, 1	13 × 13 × 384
Conv	3 × 3, 1	13 × 13 × 256
Max-pooling	3 × 3, 2	6 × 6 × 256
FC6	-	4096
FC7	-	4096
FC8, FC9	-	2, 4

(PCA) [73] of the image features for all patches within a WSI due to the correlations among the patches.

Convolution Neural Network based Features

In the recent years, with the advances of deep learning, studies using Convolutional Neural Networks (CNN) have demonstrated significant improvement on histopathology image classification [74, 75, 76, 77, 78] and segmentation [79, 80, 75, 74]. For the WSIs, applications based on CNNs are also widely developed [81, 82, 83]. In our study, we adopt two approaches to get the CNN based features. The first one is using the neural network to get image features for each patch, then the features for WSIs are obtained by utilizing PCA on all patches. The convolutional neural network employed in the study is shown in Table 3.2. The input to the network is the cropped patches from prostate pathology WSIs, and the activations from the second to the last layer are considered as the image features of the input samples. In order to train the network with patches, we assign Gleason pattern as the ground truth annotation for the patch. The GDC WSIs have been previously graded, with the primary and secondary patterns, as well as the

final Gleason score given. To model variations among Gleason patterns within a WSI, we use the multi-task architecture to enable the network to learn as much information about the Gleason patterns from the patches of a WSI as possible. During the training process, we give the primary pattern and the sum score as labels for each patch and use the following multi-task loss function:

$$\mathcal{L}_{\text{multi-task}} = -\sum_{i=0}^{N} \mathbf{t}_i^p \cdot \log \hat{\mathbf{t}}_i^p - \sum_{i=0}^{N} \mathbf{t}_i^s \cdot \log \hat{\mathbf{t}}_i^s \tag{3.1}$$

where for the i_{th} image within the batch of N images, $\mathbf{t}_i^p$ and $\mathbf{t}_i^s$ are respectively the one-hot encoding of the Gleason grading for the primary pattern and the sum score, $\hat{\mathbf{t}}_i^p$ and $\hat{\mathbf{t}}_i^s$ are respectively the predicted grading of the model. The results suggested that using the primary Gleason pattern and the Gleason score together achieved the best estimate of risk of recurrence by capturing local and global image feature distribution more efficiently than using either one alone.

For the second approach, we treated the cropped patches from the WSI as an image sequence and used one type of recurrent neural network called long-short-term memory (LSTM) to explore the long-term dynamic information of the patches spatial sequence within the WSI. We denoted the method as CNN features with LSTM (CNN+LSTM). The LSTM could fully leverage the patch spatial sequence within a WSI to get the representative features that model the global Gleason score of the WSI and the distribution of the Gleason patterns among the WSI. Recently, the LSTM model has been successfully used in speech recognition [84, 85], language translation models [86], image captioning [87] and video classification [88]. Compared with the traditional RNNs, LSTM is more effectively in long range sequence modeling. In general, given an input feature sequence (x_1, x_2, ..., x_T), the LSTM computes the output sequence as (y_1, y_2, ... , y_T). The hidden layer of LSTM is computed recursively from $t = 1$ to $t = T$ with the following equations:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{3.2}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \tag{3.3}$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{3.4}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{3.5}$$

$$h_t = o_t \tanh(c_t) \tag{3.6}$$

where x_i is the network activations of the i_{th} patch, h_t is the hidden vector, i_t, c_t, f_t and o_t are respectively the activation vectors of the input gate, memory cell, forget gate and output gate. W terms denote the weight matrices connecting different units, b terms denote the bias vectors and σ is the logistic sigmoid function. From the above equations, we can see the memory cell c_i in LSTM has two inputs: the weighted sum of the current inputs and the previous memory cell unit c_{t-1}, which enables the model to learn when to forget the old information and when to consider the new information. The output gate o_t controls the propagation of information to the following step.

Because we utilized the spatial characteristic encoded features from CNN, the training process of LSTM of patches within WSIs was formed in a spatial format instead of time sequential manner. As shown in Figure 3.3, we use the image coordinates to indicate the location of each patch in the patch spatial sequence. In this way, we consider both the unique characteristics of each patch and the fine-grained variations between patches. For one prostate WSI, the patches are fed into the network to get the activations from the second to the last layer. Then we utilize a one layer LSTM to recursively map the activations of each patch to a feature vector. In addition, the average pooling layer is applied on top of the network to get the a feature vector as the computational image features for the WSI. The number of hidden units for each LSTM is 1024. During the training process, we apply the multi-task loss and assign the primary pattern and the Gleason score for the WSIs.

3.3.2 Survival Models

To evaluate performance of various survival models using different image features quantified by textural and CNN-based methods on patients with prostate cancer, we used the bRFS since their initial treatment as a time-to-recurrence variable for survival models. Using survival models, we assess the image features related to recurrence hazard risk scores in the context of other clinical prognostic factors, including the primary and the secondary Gleason patterns, PSA, age, and clinical tumor stage.

The hazard risk scores of image features in the context of clinical mean a measure of on prostate cancer recurrence risk ratio, commonly in time-to-event analysis or survival analysis. The survival models tested in our study include multivariate Cox proportional hazards model [89], Cox regression by an elastic net penalty (COX-EN) [90], parametric proportional hazard model (PH-EX) [91], parametric proportional hazard model with log normal distance (PH-LogN) [91] and parametric proportional hazard model with log logistic distance (PH-LogL) [91].

For the high-dimensional data, univariate Cox regression is applied on the computational image features. Only those with Wald test p-value less than 0.05 are selected in conjunction with clinical factors as inputs of the survival models.

The Cox proportional hazards model is a popular regression model for analysis of survival data. It is a semi-parametric method for adjusting survival rate estimates to quantify the effect of predictor variables. In contrast with parametric models, it makes no assumptions about the shape of the so-called baseline hazard function. It represents the effects of explanatory variables as a multiplier of a common baseline hazard function H_0. Given the patients (t_i, l_i, X_i), where $i = 1, 2, ..., N$, we have the t_i as the patient's recurrence time for individual i; l_i as the label of the censored data that equals 1 if the recurrence occurred at that time and 0 if the patient has been censored; X_i as the vector of covariates of the selected image features and clinical factors.

The hazard function is the nonparametric part of the Cox proportional hazards regression function corresponding to

$$H(X_i, l_i, t_i) = H_0(t)\exp\sum_{j=1}^{p} x_{ij}\beta_i \tag{3.7}$$

Here x_{ij} is the image features j for patient i, where $j = 1, 2, ...p$ and β_i is the Cox regression parameter for each patient.

The hazard ratio is derived from $HR(X_i) = \frac{H(X_i,l_i,t)}{H_0}$, representing the relative risk of instant failure for patients having the predictive value X_i compared to the ones having the baseline values. Here d_i is weighting parameters for each patient.

$$HR(X_i) = \sum_{i}^{N} d_i(X_i\beta_i - \log(\sum_{j}^{p} I(t_j - t_i)\exp(X_i\beta_i))) \tag{3.8}$$

For the Cox regression by an elastic net penalty (COX-EN), the elastic net penalty $\hat{\beta}$ is given as below equation. It is a mixture of the L1 (Lasso) and L2 (ridge regression) penalty. Here α is the ratio between L1 and L2 for elastic net.

$$\hat{\beta} = \text{argmax}_\beta \left[\frac{2}{n} \left(\sum_{i=1}^{N} x_{i(j)}^{\mathrm{T}}\beta_i - \log(\sum_{j\in R_4} e^{x_j^{\mathrm{T}}\beta_i}) \right) - \lambda P_\alpha^{\beta_i} \right] \tag{3.9}$$

where

$$\lambda P_\alpha(\beta_i) = \lambda(\alpha\sum_{i=1}^{p}|\beta_i| + \frac{1}{2}(1-\alpha)\sum_{i=1}^{p}\beta_i^2) \tag{3.10}$$

Base on the assumption that the effect of the covariates is to increase or decrease the hazard by a proportionate amount at all durations, the parametric proportional hazard model is a location-scale model for arbitrary transform of the time variable t_i, leading to accelerated failure time model with different penalty distance functions. The distance functions we use for parametric proportional hazard models are exponential transformation (PH-EX), log normal (PH-LogN) and log logistic (PH-LogL) distances.

The survival model performance for different image feature methods were quantified by Akaike Information Criterion (AIC) [92].

$$\text{AIC} = -2\log(likelihood) + 2K \tag{3.11}$$

where *likelihood* is a measure of model fitness and K represents the number of model parameters The smaller value of the AIC, the better goodness fit of survival models.

3.4 Experimental Results

In this section, we conducted the experiments on the public prostate cancer dataset to make statistical analysis on various survival models using different histopathology image feature quantification methods.

3.4.1 Implementation Details

For the CNN based approaches to extract image features, we first use the patches to train the CNN with multi-task loss. Each patch is resized as 256×256 and assigned two labels according to the Gleason score of the WSI, one is the primary pattern and another is the Gleason score. The CNN is trained with mini-batch stochastic gradient descent. The momentum is 0.9 and weight decay is 5×10^{-5}. The initial learning rate is 10^{-3} and annealed by 0.1 after 10^4 iterations. To train the LSTM, we set the same momentum, weight decay and the initial learning rate. The learning rate is annealed by 0.1 after 2×10^3 iterations. The implementation is based on the Caffe toolbox [59].

3.4.2 Comparison of Image Features

First, only using image features from tissue specimens, including clinical Gleason primary and secondary patterns and the quantified image features from various image image methods, their Cox hazard ratios are calculated and results are shown in Table 3.3. CNN achieves better results than texture methods including SURF [69], HOG [70], and LBP [71] Using CNN with LSTM to model the spatial relation of patches achieves the highest Cox hazard ratio, which indicates the best performance on progression correlation for recurrence data. Meanwhile the image features obtained from texture based methods and CNN approaches achieve a higher Cox hazard ratios compared to utilizing primary and secondary patterns alone.

Second, in addition to the image features, PSA levels, age, and clinical tumor stage are included in the Cox survival model besides the primary and the secondary Gleason patterns. The results of combining clinical factors and image features are shown in Table 3.4 demonstrate the image features generated from CNN based approaches are

Table 3.3: The Cox hazard ratios of only using clinical Gleason primary and secondary patterns and image features from different image analysis methods. The texture feature quantificaiton methods include SURF [69], HOG [70], and LBP [71]. Using CNN with LSTM to model the spatial relation of patches achieves the highest Cox hazard ratio, which indicates the best performance on progression prediction for the recurrence data. Meanwhile the image features from texture and CNN approaches achieve the higher Cox hazard ratios compared to the ones from clinical Gleason primary and secondary patterns.

Methods	Primary Pattern	Secondary Pattern	Image Features
SURF	0.76	0.58	1.15
HOG	0.84	0.55	1.09
LBP	0.77	0.60	1.10
CNN	0.80	0.73	1.83
CNN + LSTM	0.90	0.71	3.54

Table 3.4: The Cox hazard ratios and AICs of using clinical factors including Gleason primary and secondary patterns, patient's PSA, age and clinical tumor stages and image features from different image analysis methods. The texture feature quantification methods include SURF[69], HOG[70], and LBP[71]. Using CNN+LSTM to achieves the highest Cox hazard ratio and lowest value of AIC, which indicates the best performance on progression prediction for the recurrence data.

Methods	Primary Pattern	Sec-ondary Pattern	PSA	Age	Tumor Stage	Image Features	AIC
SURF	0.99	0.67	0.84	0.98	1.04	1.13	38.93
HOG	1.21	0.65	0.82	1.01	1.13	1.10	51.97
LBP	0.97	0.76	0.84	1.00	1.08	1.08	35.97
CNN	1.10	1.13	0.80	1.00	1.17	2.58	38.02
CNN + LSTM	1.38	0.75	0.76	0.97	1.14	**7.10**	**35.60**

Table 3.5: The Cox hazard ratios of the clinical factors.

Primary Pattern	Secondary Pattern	PSA	Age	Tumor Stage
2.15	1.09	0.73	0.9	1.3

Table 3.6: The hazard ratios and AICs of CNN-based approaches on patient progression analysis using three different training strategies. Using multi-task architecture achieves the highest Cox hazard ratio and lowest AIC values than training using the primary Gleason pattern or Gleason score alone, which indicates the best performance on progression prediction for the recurrence data.

Methods	Training Strategy	Primary Pattern	Secondary Pattern	PSA	Age	Tumor Stage	Image Features	AIC
CNN	Primary Pattern	1.11	1.12	0.80	1.00	1.16	1.34	46.13
CNN	Gleason Score	1.26	1.03	0.75	0.98	1.12	1.53	44.29
CNN	Multi-task	1.10	1.13	0.80	1.00	1.17	2.58	38.02
CNN + LSTM	Primary Pattern	1.35	0.84	0.78	0.98	1.14	1.63	44.27
CNN + LSTM	Gleason Score	1.09	0.66	0.81	0.99	1.11	2.76	41.47
CNN + LSTM	Multi-task	1.38	0.75	0.76	0.97	1.14	7.10	35.60

more representative than the texture features by having higher values of hazard ratio. Additionally, those features are more representative than the clinical prognostic factors. We also show the AIC values in Table 3.4, from which we can see CNN+LSTM achieves the best fitness on the Cox regression model.

Finally, without any image features, we show the Cox hazard ratios of the clinical factors in Table 3.5. From the results in Table 3.5, Table 3.3, and Table 3.4, we can see primary Gleason patterns has higher Cox hazard ratios than the ones of other clinical factors, which is consistent with its high prediction power for prostate cancers [63, 64]. Besides image-related clinical factors, in conjugation with other clinical factors, such as patient's PSA, age and clinical tumor stage could increase the Cox hazard ratios of image features.

3.4.3 Ablation Study on Training Strategies

Furthermore, considering the multiple Gleason patterns within WSIs, we design two training strategies to train the CNN-based approaches. The first one is to use multi-task loss to learn both the primary Gleason pattern and the sum of the primary and secondary patterns (namely, the Gleason score). The second one is to use the primary Gleason pattern or the Gleason score alone to learn the patterns within the patches or WSIs.

The performance of two CNN-based approaches on patient survival analysis are compared using different training strategies. The results are shown as in Table 3.6. We can see the multi-task architecture achieves better survival performance than training label using the primary Gleason pattern or Gleason score alone as it has much higher recurrence hazard ratios and lower AIC values. Because the primary Gleason pattern and the Gleason score together could better reflect the local and global image features in the WSIs than use each alone.

3.4.4 Comparison of Survival Models

In this section, we do the statistical analysis on various survival models, including COX-EN [90], PH-EN [91], PH-LogN [91], and PH-LogL [91], using prostate images with Gleason score 6 to 8 and clinical factors. The Cox proportional hazards model does not need an assumption of a particular survival distribution of the patients' survival data. The only assumption in the mode is about the proportional hazards. Unlike the Cox proportional hazards model, the parametric model with different penalty distance functions (such as exponential, log-normal and log-logistic) need to specify the hazard functions [93, 94]. Studies have indicated that under certain circumstances, such as strong effect or strong time trend in covariates or follow-up depending on covariates, parametric models are good alternatives to the Coxs regression model [94].

We assess different survival models and show the hazard ratios of image features and patients' clinical prognostic factors in Table 3.7. Based on these results, first, we can see the image features quantified from WSIs outperform other clinical factors in all

Table 3.7: Hazard ratios and AICs of different survival models using texture methods and CNN-based approaches. The survival models include COX-EN [90], PH-EX [91], PH-LogN [91], and PH-LogL [91].

Survival Models	Methods	Primary Pattern	Secondary Pattern	PSA	Age	Tumor Stage	Image Features	AIC
COX-EN	SURF	0.10	0.27	0.33	0.06	0.03	3.38	42.93
COX-EN	HOG	0.10	0.25	0.32	0.06	0.03	3.85	59.72
COX-EN	LBP	0.10	0.19	0.30	0.06	0.03	2.40	39.83
COX-EN	CNN	0.23	0.21	0.33	0.06	0.04	7.57	29.86
COX-EN	CNN + LSTM	0.13	0.27	0.36	0.06	0.03	**15.85**	**29.83**
PH-EX	SURF	0.07	0.09	0.29	0.03	0.03	1.94	41.26
PH-EX	HOG	0.05	0.12	0.29	0.04	0.03	2.41	61.56
PH-EX	LBP	0.07	0.06	0.28	0.03	0.03	1.49	41.22
PH-EX	CNN	0.08	0.07	0.29	0.04	0.04	4.50	35.60
PH-EX	CNN + LSTM	0.08	0.10	0.29	0.04	0.03	10.22	31.22
PH-LogN	SURF	0.18	0.22	0.30	0.02	0.08	2.03	47.27
PH-LogN	HOG	0.18	0.23	0.30	0.02	0.08	2.70	47.58
PH-LogN	LBP	0.21	0.18	0.29	0.02	0.08	1.38	45.99
PH-LogN	CNN	0.16	0.15	0.30	0.02	0.08	4.33	42.51
PH-LogN	CNN + LSTM	0.20	0.18	0.31	0.02	0.08	11.92	33.31
PH-LogL	SURF	0.11	0.15	0.29	0.02	0.07	1.89	43.74
PH-LogL	HOG	0.07	0.20	0.28	0.02	0.06	2.91	44.45
PH-LogL	LBP	0.79	0.29	1.09	0.77	0.18	1.46	44.39
PH-LogL	CNN	0.09	0.08	0.29	0.03	0.07	4.39	35.96
PH-LogL	CNN + LSTM	0.12	0.13	0.29	0.02	0.07	9.92	33.02

texture and CNN-based approaches. Second, CNN-based approaches achieve a better progression prediction due to their higher hazard ratios than other texture methods for all survival models. Third, by comparing with Table 3.4, COX-EN achieves the lowest AIC value with image features obtained from CNN+LSTM, proving that the model is more suitable for recurrence analysis for prostate patients with low, intermediate, and high risk than other survival models.

3.5 Discussion and Conclusion

In this paper, we present three texture methods (SURF, HOG and LBP) and two convolution neural network (CNN) based methods to quantify features from histopathology images. Five survival models were assessed on those image features in the context with prostate clinical prognostic factors including the primary and the secondary Gleason patterns, PSA, age and clinical tumor stage to perform recurrence analysis for prostate cancer.

From statistical comparisons among different image feature quantification methods with survival models, the CNN-LSTM provided the highest hazard ratio of prostate cancer recurrence under Cox regression by an elastic net penalty (COX-EN). It outperforms other image quantification methods with other survival models respectively. From our approach, patient outcomes were better correlated with their histopathology images. Due to the limited size of the public prostate dataset, the results achieved from our experiments are preliminary. In order to further validate its generalizability of our approach, more prostate images from local institutions are needed to perform extensive experiments.

Chapter 4

Recurrence Analysis on Prostate Cancer Patients with Gleason Score 7 using Integrated Histopathology Whole-Slide Images and Genomic Data through Deep Neural Networks

4.1 Introduction

Prostate cancer remains the most common non-cutaneous malignant tumor in the western world accounting for approximately 1 in 5 of newly diagnosed tumors in men [1]. In the United States, approximately 1 in 7 men will be diagnosed with this disease [1]. Based on Gleason score, prostate specific antigen (PSA) value, tumor stage, age and race, patients with prostate cancer are stratified into low-risk, intermediate-risk and high-risk groups [95].

A strong predictor of survival among men with prostate cancer is the Gleason score that is rendered by a pathologist based upon a microscopic evaluation of a representative histopathology specimen [96]. These scores are based solely upon morphology and structural patterns of the constituent cells and glands. Patients with Gleason score 6 or lower often undergo active surveillance as there is reduced risk of tumor progression for those patients compared to patients with score 7 or higher[97, 98]. Tumors that are assigned Gleason score 7 can be delineated into a primary region exhibiting a histopathology pattern graded as 4 and a secondary region exhibiting a histopathology pattern graded as 3. Such samples are referred to as Gleason 4+3 tumors, whereas the inverse pattern exhibiting a primary pattern of 3 and a secondary pattern of 4 would constitute a Gleason 3+4 tumors. Patients with Gleason 4+3 tumors have an increased risk of recurrence and progression leading to an increased risk of prostate cancer specific mortality when compared to those afflicted with Gleason 3+4 tumors [99, 100, 101].

The literature clearly shows that predicting disease recurrence in a man with Gleason score 7 prostate cancer can have a significant impact on a his disease management and survival [100, 101, 17].

Phenotypically, tumor regions with Gleason pattern 3 are composed of single glands with distinct size and shape whereas ones with Gleason pattern 4 exhibit large irregular cribriform glands or fused, ill-defined glands with poorly formed glandular lumina [102]. In spite of established guidelines, Gleason grading remains a relatively subjective process that results in an approximately 30% grading discrepancy among the scores provided by pathologists [102]. There have been many attempts to develop computer-aided Gleason grading methods and systems [103, 52, 102, 50, 25, 26, 104] in order to introduce objective, reproducible criteria into the process of Gleason pattern quantification and grading. One previous study has explored an integration of image features along with protein expression to predict recurrent prostate cancer [61]. However, to date there has been no study focused on utilizing patients' pathology images and genomic pathway analyses in combination to predict recurrence-free survival (RFS) for men with prostate cancer.

Microarray-based gene expression signatures have been used in various studies to identify cancer subtypes, determine the RFS of disease and and characterize response to specific therapies [105]. Multiple investigations have also shown that gene expression signatures can be used to analyze oncogenic pathways and these signatures have been used to identify differences between specific cancer types and tumor subtypes. Moreover, patterns of oncogenic pathway activity have been used to identify differences in underlying molecular mechanisms and have been shown to correlate with clinical outcomes of patients afflicted with specific cancers [106, 107, 108, 42].

In recent years, whole-slide image (WSI) has been more widely used in histopathology diagnosis. With a fast development of deep learning, histopathology image analysis approaches have demonstrated significant advances in cellular segmentation [79, 80, 75, 74] and tissue classifications [74, 75, 76, 77, 78] using Convolutional Neural Networks (CNN). Some research groups reported their studies using histopathology WSI for many applications [81, 82, 83]. Due to a giga-pixel size of a WSI's, it is often impractical to

train the CNN using WSIs directly. Consequently, patch-based algorithms are widely applied in histopathology image analysis [27, 28, 29, 30, 31, 46].

In this study, we developed a computational biomarker quantification system by integrating histopathology WSIs and genomic data into one deep neural network. In order to use the distribution of Gleason patterns on a WSI, we applied patches as inputs to the network. The patches were forwarded through a CNN to get the images features. Then based on the patches' spatial relationship, the image features were modeled using a recurrence neural network (RNN) [40], namely long short-term memory (LSTM) [41]. The pathway scores calculated from the genomic data were forwarded to a multilayer perceptron (MLP) to get the genomic features. And the image and genomic features were integrated together to get the computational biomarkers. Moreover, we used RFS (months) since their initial treatment as the time-to-recurrence variable for a survival model. We chose a Cox proportional-hazard regression model [94, 109], since it is commonly used in medical research for investigating associations between the survival time of patients and predictor variables.

4.2 Methods

In this section, we introduced our approach on building a unified system using WSI and genomic data through deep neural networks to quantify computational biomarkers, which were fed into a survival model for patients' recurrence analysis. Our methods consisted of four steps. Firstly, the pathway activities of prostate cancer were quantified by pathway scores using RNA sequences. Secondly, the histopathology WSIs were pre-processed to obtain the region-of-interest (ROI) as the image patches preparation. Thirdly, the image patches and pathway scores were integrated into a unified system using the deep learning approach to extract computational biomarkers. Finally, we used the computational biomarkers in conjunction with clinical prognostic factors as the input of the survival model to calculate the disease recurrence ratios and probabilities. Figure 4.1 illustrates the overview of the pipeline of the whole study.

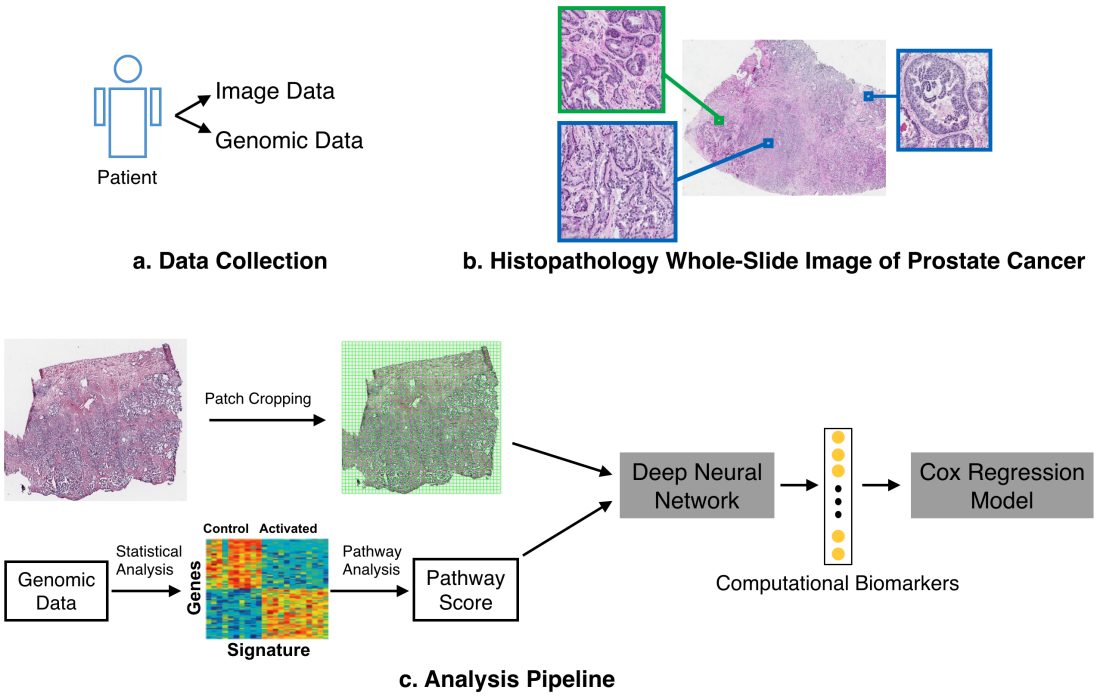


Figure 4.1: An overview of the pipeline of our study using histopathology WSIs and genomic data for prostate cancer recurrence prediction for patients with Gleason score 7. (a) WSI images and genomic data were collected from patients with prostate cancer; (b) A prostate WSI exhibits different Gleason patterns. For example, a region in a green square has the Gleason pattern 3 while regions in blue squares have the Gleason pattern 4; (c) The pathway scores were quantified using RNA sequences. Patches of region of interests were automatically selected from WSIs. The image patches and pathway scores were integrated into deep neural networks to extract computational biomarkers, which were fed into a Cox regression model in conjunction with clinical prognostic factors for disease recurrence analysis.

4.2.1 Experiment dataset

In this study, we used publicly available prostate cancer data downloaded from the data portal[1] of the Genomic Data Commons (GDC) [110]. GDC is the largest public available data portal that includes image data from The Cancer Genome Atlas (TCGA) [67], genomic data and clinical data. The TCGA barcode[2] is the primary identifier of GDC data acquisition protocol. For this study, in total, there were 43 Gleason 3+3, 146 Gleason 3+4, 101 Gleason 4+3 and 49 Gleason 4+4, which contains 1229, 4753, 2997 and 1597 patches respectively. For the recurrence study of patients with Gleason 7, we used all the data from Gleason 6, 7 and 8 to train the networks to extract the computational biomarkers. In this way, the training data contained more images of Gleason patterns 3 and 4 compared to a training data if only use patients' data with Gleason 7 (3+4 or 4+3). For the recurrence study of patients with Gleason 7, the computational biomarkers of patients with Gleason 7 were fed into a survival model, while the patients with other Gleason score were withheld.

The patients were randomly divided into the training set, validation set and testing test with the ratio of 70%, 10% and 20%; these groups were utilized for the recurrence analyses. In addition to the Gleason score, we compared the computational biomarkers quantified from the unified image and genomic data system with other clinical factors including patients' PSA, age and tumor stage, which are publicly available from GDC data portal.

The WSI patches preparation was a two-step cropping-selection process. Firstly, the image patches within each WSI were automatically cropped under 40× objective magnification with a patch size 4096×4096. The patches were cropped with a stride as 4096 to avoid overlapping. We resized all the patches to the size of 256×256 using Lanczos filtering [111]. Secondly, any specimens with insufficient tissue patches were automatically eliminated from the experiments due to the heterogeneous quality of the prostate WSIs. The patches with the tissue accounting for at least 20% of the whole

[1] https://portal.gdc.cancer.gov

[2] https://docs.gdc.cancer.gov/Encyclopedia/pages/images/TCGA-TCGAbarcode-080518-1750-4378.pdf

area were selected.

4.2.2 Pathway scores quantification from RNA sequencing data

To quantify pathway scores, we used the gene expression data, which were RNA (Illumina HiSeq) sequencing data from patients with Gleason score 7. The data are publicly available through GDC data portal. We preprocessed the RNA data by log transformed and median centered. A panel of previously published 265 experimentally derived gene expression signatures were applied to the entire cohort to identify patterns of oncogenic signaling in each tumor. To apply a given signature, the expression data were filtered to contain only those genes included in the given signature and the mean expression value of these genes was calculated to provide a score for each sample [107, 108] .

4.2.3 Computational biomarkers extraction

In order to obtain computational biomakers from the WSIs and genomic data, we built a unified feature quantification system using CNN to model WSI histopathology image patches and genomic data together. Furthermore, we leveraged the RNN to model the spatial relationship of the cropped patches within the WSI. The network architecture is shown in Figure 4.2.

Modeling histopathology image patches and genomic data

In order to combine the image information along with the genomic data, we used the patches and pathway scores as the input to the network. We forwarded the pathway scores into a multilayer perceptron that includes three fully connected (FC) layers, with 1024, 512 and 256 hidden units, respectively. The genomic features were the output of the last FC layer. Meanwhile, we incorporated the AlexNet [112] to extract the features from image patches. We concatenated the genomic features obtained from the pathway scores with the image features from the second to the last layer of the AlexNet. The concatenated features were served as the input to the a FC layer before LSTM.

Due to the giga-pixel WSI's, we considered an integrity of the whole tissue regions on a single WSI instead of using the individual patches to quantify image features as

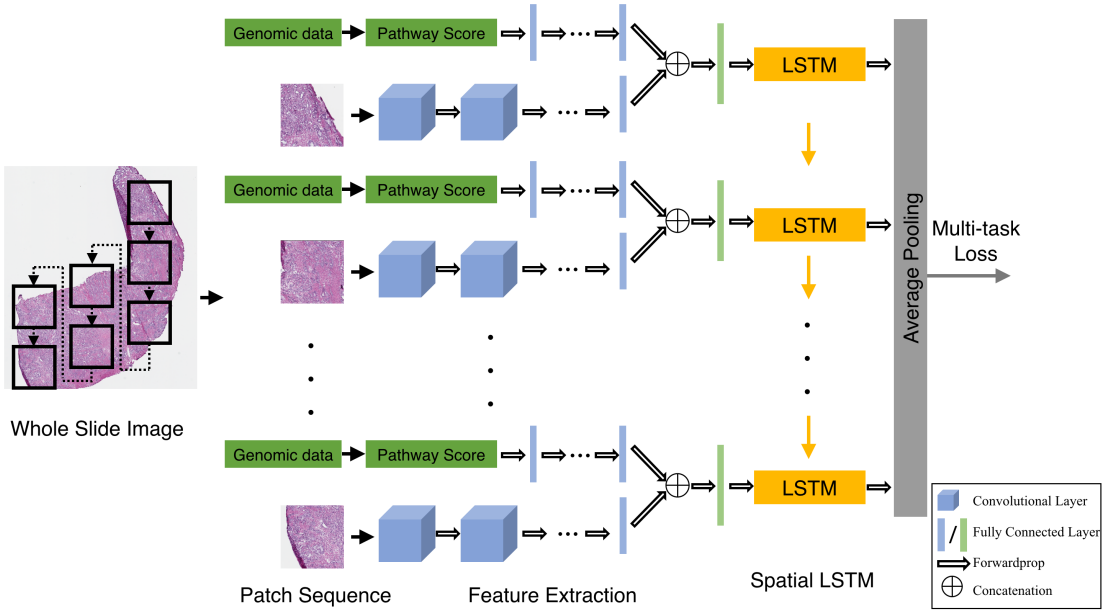


Figure 4.2: Network architecture for extracting computational biomarkers from the WSI and genomic data. We used seven LSTM cells in the network. The calculated pathway scores from the genomic data were forwarded into a multilayer perceptron (MLP) that contains three fully connected layers. The last layer of the MLP was connected with the features extracted from the image patches to serve as the input for the LSTM after a fully connected layer. On top of the LSTM, we utilized an average pooling layer.

shown in previous studies. [30, 28] The spatial relationship of the adjacent patches was modeled as an image sequence. We adopted a type of recurrent neural network (RNN) [40], long short-term memory (LSTM) [41], to model the features extracted from the image patches and genomic data given LSTM has shown its successes among various applications including speech recognition [84, 85], language translation models [86], image captioning [87] and video classification [88]. Compared with the traditional RNN that has vanishing and exploding gradients problems, LSTM is more effectively in sequence modeling by incorporating memory cells with several gates to obtain long-range dependencies.

More formally, for the input feature sequence (x_1, x_2, ..., x_T) that x_i represents the activations from the CNN of the i_{th} patch, we used LSTM to compute the output sequence (y_1, y_2, ... , y_T), where the layer layer of LSTM was computed recursively from $t = 1$ to $t = T$ following the equations:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{4.1}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \tag{4.2}$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{4.3}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{4.4}$$

$$h_t = o_t \tanh(c_t) \tag{4.5}$$

where h_t is the hidden vector, i_t, c_t, f_t and o_t represents the activation vectors of the input gate, memory cell, forget gate and output gate, respectively. W terms denote the weight matrices connecting different units, b terms denote the bias vectors and σ is the logistic sigmoid function. The memory cell c_i has the inputs of the the weighted sum of the current inputs and the previous memory cell unit c_{t-1}, which could learn when

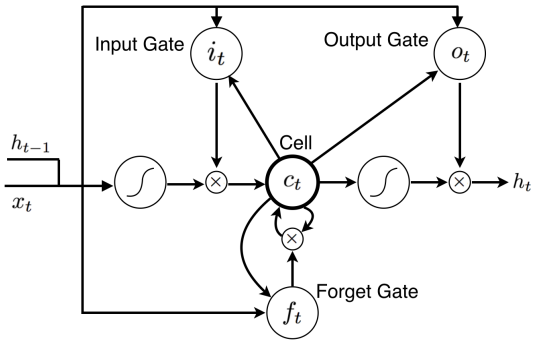


Figure 4.3: The visualization of a LSTM cell.

to forget the old information and when to consider the new information. The output gate o_t controls the propagation of information to the following step. The visualization of the LSTM cell is shown in Figure 4.3.

Since it is a sequential task to train LSTM, patches from a WSI were formed by a specific routine. As demonstrated in Figure 4.2, we used center coordinates of patches to remark the location of each patch. The sequence of patches within a single WSI was arranged from right up patch down to lower left one, which was illustrated by the dotted lines with black arrows on an example of a WSI on Figure 4.2. In this way, it allowed us to consider both unique characteristics of each patch and fine-grained variations among patches within a single WSI. For each tumor WSI, the patches and the pathway scores were fed into the network to get features and then incorporated into the LSTM recursively. In addition, the average pooling layer was applied on top of the network to get the computational biomarkers for the WSI and the genomic data. The number of hidden units for each LSTM was 1024. During the training process, we applied the multi-tasks loss and assigned the primary pattern and the Gleason score for the WSIs and genomic data.

Multi-tasks loss function

For the TCGA prostate WSIs, the primary Gleason pattern, the secondary Gleason pattern and the sum of both patterns (i.e. Gleason score) were publicly available from

GDC data portal. To model the variations among Gleason patterns, we utilized the multi-task loss to enable the network to learn as much information about the Gleason pattern distributions from the patches of a WSI as possible. Therefore, we gave the primary pattern and the sum score as labels for each patch along with the pathway score and use the following multi-task loss function:

$$\mathcal{L}_{\text{multi-task}} = -\sum_{i=0}^{N} \mathbf{t}_i^p \cdot \log \hat{\mathbf{t}}_i^p - \sum_{i=0}^{N} \mathbf{t}_i^s \cdot \log \hat{\mathbf{t}}_i^s \tag{4.6}$$

where for the i_{th} input sample within the batch of N samples, $\mathbf{t}_i^p$ and $\mathbf{t}_i^s$ are respectively the one-hot encoding of the Gleason grading for the primary pattern and the sum score, $\hat{\mathbf{t}}_i^p$ and $\hat{\mathbf{t}}_i^s$ are respectively the predicted grading of the model.

4.2.4 Survival model

In conjunction with clinical prognostic factors including the primary and secondary Gleason patterns, PSA, age and tumor stage, computational biomarkers were fed into a Cox regression model [94, 109] for studying patients RFS. In our study, we used RFS (months) as the time variable for a survival model. For high dimensional data, only those with Wald test[113, 114] p-value < 0.05 were selected and used in conjunction with clinical prognostic factors as input variables for the Cox regression model.

One of the most popular regression techniques for survival analysis is Cox proportional hazards regression, which is used to relate several risk factors or exposures, considered simultaneously, to assess differences in overall survival. In a Cox proportional hazards regression model, the measure of effect is the hazard ratio, which is the risk of failure (i.e., here is the risk or probability of the recurrence of the disease), given that the participant has survived up to a specific time. Given the patients (t_i, l_i, X_i), where $i = 1, 2, ..., N$, we have the t_i as the patient's recurrence time for individual i; l_i as the label of the censored data that equals 1 if the recurrence occurred at that time and 0 if the patient has been censored; X_i as the vector of covariates of the selected image features and clinical factors. The hazard function is the nonparametric part of

the Cox proportional hazards regression function corresponding to

$$H(X_i, l_i, t_i) = H_0(t)\exp\sum_{j=1}^{p} x_{ij}\beta_i \tag{4.7}$$

Here x_{ij} is the computational biomarkers j for patient i, where $j = 1, 2, ...p$ and β_i is the Cox regression parameter for each patient. Here H_0 is the baseline hazard function. The hazard ratio is derived from $HR(X_i) = \frac{H(X_i,l_i,t)}{H_0}$, representing the relative risk of instant failure for patients having the predictive value X_i compared to the ones having the baseline values. Here d_i is weighting parameters for each patient.

$$HR(X_i) = \sum_{i}^{N} d_i(X_i\beta_i - \log(\sum_{j}^{p} I(t_j - t_i)\exp(X_i\beta_i))) \tag{4.8}$$

In the study, we assessed the computational biomarkers in conjunction with other clinical prognostic factors by their recurrence hazard ratios and concordance indices (C-index)[115, 116]. The hazard ratio and C-index both are global indices for validating the predictive ability of prognostic features of a given survival model. Under a given survival model, higher values mean that prognostic features predict higher risks and probabilities of survival for higher observed survival times. In our study we examined RFS; the higher the hazard ratio and C-index, the greater the likelihood of disease recurrence.

4.3 Experiments and Results

In this section, we validated our approach on a publicly available prostate cancer dataset from the GDC data portal. The experimental results showed the computational biomarkers discovered by the proposed method were effective for recurrence correlation for patients with Gleason score 7.

4.3.1 Implementation details

The training process of our network included two steps. We first trained the CNN using mini-batch Stochastic Gradient Descent (SGD) with batch size as 32, momentum as 0.9, and weight-decay as 5×10^{-5}. The initial learning rate was 10^{-3} and annealed by

0.1 after every 10, 000 iterations. We trained the CNN for total of 50, 000 iterations until the loss converge. Then we utilized the genomic data to train the MLP to extract the genomic features and used image and genomic features to train LSTM. We kept the same momentum, weight-decay and learning rate except we annealed the learning rate by 0.1 after every 2, 000 iterations and trained the network for a total of 5, 000 iterations. The implementation was based on Caffe toolbox[59].

4.3.2 Pathway analysis

Multiple studies have shown that gene expression signatures reflect the activation status of oncogenic pathways irrespective of specific mutations driving signaling[106, 108, 107]. Thus we examined genomic-based patterns of oncogenic pathway activity in prostate cancer patients with Gleason score 7 using a panel of previously published 265 gene expression signatures.

In order to qualitatively assess unique patterns of pathway activity that define the 4+3 and 3+4 subset of Gleason score 7 tumors, pathway signatures in each group, using all tumors across the entire cohort (i.e. training, test and validation tumors) were assessed by a Student's two tailed t-test. Significant pathway scores were clustered using Cluster 3.0 [117] and visualized by Java TreeView [118]. Quantitative assessment of patterns of pathway activity of Gleason score 4+3 and 3+4 subgroups is shown in Figure 4.4, which displayed a heatmap identifying 27 differentially expressed signatures ($p<0.01$). Of these, we determined that 14 signatures including three unique proliferation signatures (Wirapati [119], UNC [120], and Murine Proliferation [120]) as well as several proliferation-associated signatures predicative of BMYB [121], RB-LOSS [122], PIK3CA [123] and HERI [124] signaling were significantly higher in patients with Gleason score 4+3. We further determined that 13 signatures were up-regulated in Gleason 3+4 patients including immune systems signatures associated with Th17 cells [125], Tcm [125], NK-CD56 [125], HGF [126] and STAT3 [108] signaling. Consistent with our findings, many studies report [62, 63, 64] that Gleason 3+4 tumors have a better prognosis than Gleason 4+3 tumors which would correlate with relatively higher levels of proliferation as well as lower levels of immune-related signaling evident in Gleason

4+3 tumors compared to Gleason 3+4 samples.

4.3.3 Integrated recurrence analysis in conjunction with clinical factors

Image data on recurrence analysis

For the integrated recurrence analysis using a survival model, we first conducted the experiments where only the WSIs of tissue slides were used. Thus the networks were trained without integrating the genomic features. This setting of experiment is denoted as CNN-LSTM. We also considered the setting that only CNN was applied on the image patches without considering their spatial relation on a WSI and the image features were extracted from the second to the last layer of AlexNet. The setting is denoted as CNN-Only. To compare the effectiveness of the feature extraction from the images, we applied three texture feature methods including SURF[69], HOG[70] and LBP[71] on the WSIs to obtain image features. The image features were concatenated with clinical prognostic factors as multivariate inputs of the Cox regression model. During each iteration, each image feature in conjunction with clinical factors were fed into the Cox regression model to calculate the corresponding hazard ratios and C-indices. The survival model implementation was based on a R survival package [127].

The maximum hazard ratios of recurrence of image features in conjunction with clinical factors are shown in Table 4.1. Within our study, we used the disease RFS times as the time variable in the Cox regression model, the higher values of hazard ratio and C-index of the features indicated that the image features had the higher correlations with the disease recurrence and progression. From the result of using texture features, there was no significance differences between LBP, HOG and SURF for recurrence ratios. CNN-LSTM analysis determined that image features identified by computational image analysis outperformed other texture features and CNN-Only with higher hazard ratio and C-index. When conjunction with CNN-LSMT, the primary pattern still showed greater hazard ratio and C-index relative to those identified using other methods.

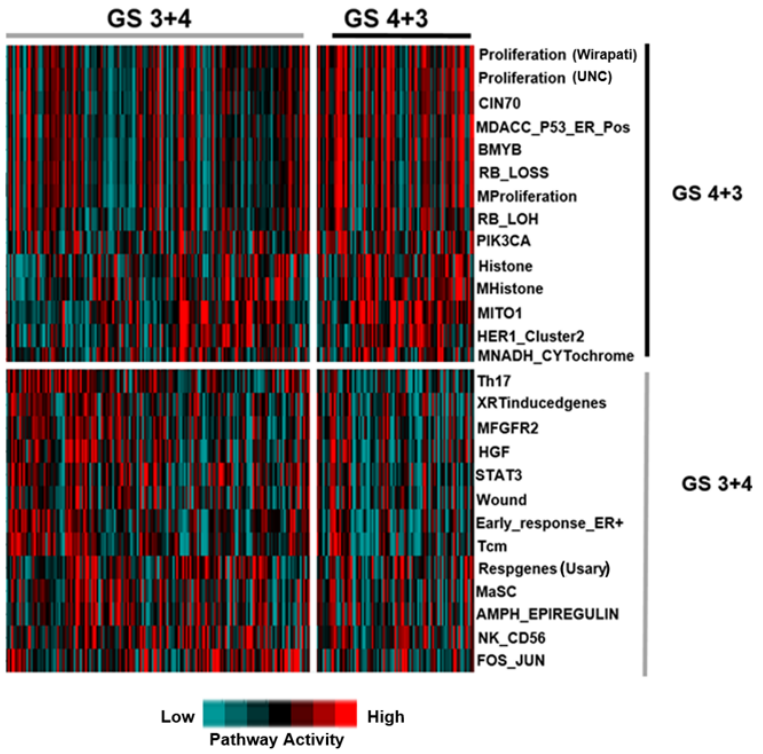


Figure 4.4: Differential patterns of pathway activity in Gleason score 3+4 and 4+3 prostate tumors. Comparative analysis of Gleason Score 4+3 (n=101) and Gleason Score 3+4 (n=146) tumors identified 27 significantly altered signaling pathways (t-test, $p<0.01$) as defined by mRNA-based gene expression signature scores. Tumors with a Gleason score 4+3 showed higher proliferation, BMYB, RB-LOH and histone modification signature scores while tumors with a Gleason score 3+4 showed higher levels of immune system related pathway signatures including Th17 cells, Tcm and STAT3.

Table 4.1: Recurrence hazard ratios and corresponding C-indices of clinical prognostic factors and different image features from various image quantification methods. The results are obtained by using image features quantified from the WSIs. LBP, HOG and SURF are the texture methods. CNN-LSTM is using the image features obtained from CNN with LSTM while CNN-Only is using the image features obtained from CNN without considering patches' spatial relation on a WSI.

Methods	Primary Pattern	Secondary Pattern	PSA	Age	Tumor Stage	Image Features	C-index
LBP	1.05	0.94	0.85	1.00	1.03	1.05	0.68
HOG	1.04	0.94	0.85	1.00	1.03	1.05	0.64
SURF	1.07	0.97	0.86	1.00	1.03	1.05	0.61
CNN-Only	1.11	1.12	0.80	1.00	1.17	2.44	0.70
CNN-LSTM	1.70	1.06	0.80	0.99	1.26	5.06	0.71

Image and genomic data on recurrence analysis

Before integrating image features and pathway scores, we first analyzed the correlation between them. Because the number of image features and the number of pathway scores were different, to calculate their correlation coefficients, we randomly chose the same number of image features paired with the same number of pathway scores and repeated the process N times until all image features had been paired. Here the image features included features quantified from texture methods (LBP, HOG, and SURF) and CNN-LSTM. Using a t-test on correlation coefficients, the mean and standard deviation of p-values is shown in Table 4.2. Because p-value > 0.05, there was no significant correlations between image features and pathway scores. This showed that the two types of data provided complementary information for prostate cancer diagnosis and prognosis. It was reasonable to integrate image and genomic data together for predicting patients' recurrence.

Then we showed the experimental results by combining image features obtained from WSIs and the genomic features obtained from the pathway scores. We utilized all 265 gene expression signatures integrated with image data to identify the computational biomarkers as shown in Figure 4.2. The setting was denoted as CNN-LSTM+PS. We also considered the setting where LSTM was deactivated when obtained the biomarkers from image and genomic data. We denoted the approach as CNN-Only+PS. The

Table 4.2: Correlation analysis of image features and pathways scores using a test-test on their correlation coefficients.

Image Features	mean of p-value	standard deviation of p-value
LBP	0.50	0.29
HOG	0.49	0.30
SURF	0.43	0.30
CNN-LSTM	0.48	0.29

methods using texture features obtained from WSIs together with pathway scores for the recurrence analysis were denoted as LBP-PS, HOG-PS and SURF-PS. We also considered only using pathway scores with clinical factors together as the input of the Cox regression model and denote it as PS. The maximum hazard ratios of the computational biomarkers from WSIs and pathway scores in conjunction with clinical factors are shown in Table 4.3.

Compared with other clinical factors, using pathway scores alone achieved equivalent hazard ratio. For the texture methods, the recurrence hazard ratios were equivalent to the ones without pathway scores. Using CNN-LSTM+PS, the Gleason primary pattern and computational biomarkers showed the increased recurrence ratios compared to the ones without pathway scores. In addition, the Gleason primary pattern and computational biomarkers showed the highest recurrence ratios compared to other clinical factors. The result showed CNN-LSTM-PS outperformed other methods in prostate cancer recurrence analysis due to its highest recurrence hazard ratio.

Furthermore, we show the C-index of the clinical factors and computational features under the Cox regression model for prostate cancer recurrence probability prediction in the last column of Table 4.1 and the last row and column of Table 4.3. As a global index for validating the predictive ability of a survival model, in our study, the C-index was equivalent to a rank correlation of the risk of a recurrence of disease. High values mean that the model predicts higher probabilities of recurrence for higher observed recurrence times. From the clinical results, PSA showed higher C-index values which were correlated to a higher recurrence prediction probability compared to other clinical

Table 4.3: Recurrence hazard ratios and corresponding C-indices of clinical prognostic factors and computational biomarkers under a Cox regression model using different image feature quantification methods along with the genomic data. Given the genomic data, we show the results using image features with pathway scores (PS). Here LBP+PS, HOG+PS, SURF+PS, CNN-Only+PS and CNN-LSTM+PS are image features quantified from LBP, HOG, SURF, CNN-Only and CNN-LSTM methods with PS.

Methods	Primary Pattern	Secondary Pattern	PSA	Age	Tumor Stage	Image Features	C-index
PS	0.95	0.98	0.86	1.00	1.04	1.02	0.65
LBP+PS	1.04	1.00	0.87	1.00	1.02	1.08	0.69
HOG+PS	1.04	1.00	0.87	1.00	1.02	1.08	0.65
SURF+PS	1.07	1.00	0.86	1.00	1.03	1.07	0.62
CNN-Only+PS	1.13	1.11	0.80	1.00	1.17	2.58	0.71
CNN-LSTM+PS	2.56	0.63	0.66	1.01	1.05	5.73	0.74
C-index for Clinical Factors	0.61	0.59	0.66	0.55	0.53	-	-

factors. Interestingly, texture features on WSIs or pathway scores individually showed an equivalent recurrence probability.

4.4 Discussion

From the experimental results, our proposed method achieved the highest recurrence hazard ratio and the strongest C-index related to prostate cancer recurrence probability compared to other clinical prognostic factors and methods. It demonstrated that the approach was beneficial for recurrence analysis on the patients with Gleason score 7. The unified WSIs and genomic data analysis through the proposed networks could be applied to other prostate cancer risk group such as Gleason 6 [128, 129, 130] or other cancer recurrence analysis, such as breast cancer [131].

Pathway analysis, albeit with the caveat of a small sample size, identified 27 differentially expressed pathway activities in tumors with Gleason score 3+4 and 4+3. Thus these signatures could be utilized to differentiate patients with Gleason score 7 as two

sub-groups which corresponds with a favorable or unfavorable prognosis [132]. While the recurrence analysis (Table 4.3), using pathway scores alone did not show an advantage over other clinical prognostic factors. The integration of pathway score with WSIs achieved the best recurrence prediction on patients with Gleason score 7. The comparison indicated using the pathway scores directly had a limited contribution in recurrence prediction on patients with Gleason score 7. However, the embedded genomic features obtained through MLP were more effective for prostate cancer recurrence analysis.

There are other clinical factors for prostate cancer prognosis besides those used in the study, such as patients race. Because in the study less than 2% men were Asian or African, 30% were Caucasian and the rest were unknown, we excluded patients race factor in the recurrence analysis. Other clinical factors, such as American joint committee on cancer metastasis stage, neoplasm disease stage codes and so on, were not available for all the patients in the GDC prostate cancer datasets.

The prostate cancer datasets were acquired from various institutions and each institution may have different scanners or WSI scanning protocols. Thus there was color heterogeneity among the prostate cancer WSIs. In the study, we did not adopt color normalization [133] on the randomly selected testing set because it was not feasible to determine the reference image from the training set for color normalization. When apply the approach to a new dataset, we could fine-tune the network based on the training data from that dataset. Given the limited size of the public prostate dataset, the results achieved from our experiments were preliminary. In order to further validate the generalizability of our approach on a wider population of prostate cancer patients, we will collect more prostate images from local institutions to perform extensive experiments.

4.5 Conclusion

In this study, we performed recurrence analyses for prostate cancer patients with Gleason score 7 integrating histopathology WSIs and genomic data. The image features and genomic features were obtained using CNN and MLP respectively. The combination of

the features were modeled using LSTM to get the computational biomarkers. Experimental results utilizing on publicly available prostate cancer dataset showed that the computational biomarkers extracted by our approach were more closely correlated with patients recurrence risk compared to standard clinical prognostic factors and engineered image texture features. The results of our study suggest that these approaches could be utilized to predict recurrence and progression for patients with prostate cancer.

Chapter 5

Factorized Adversarial Networks for Unsupervised Domain Adaptation

5.1 Introduction

Rapid development of deep convolutional neural networks (CNN) has led to promising performance on various computer vision tasks [112, 134, 135], especially with the help of large-scale annotated datasets, such as ImageNet [136]. However, when a model learned from a large dataset in one domain (source domain) is applied to another domain (target domain) with some different characteristics, it is not guaranteed to generalize well. In order to mitigate the influence caused by domain shift [137], two major approaches are widely employed. One popular approach is to fine-tune the model learned from source domain using annotated data from target distribution [32]. However, this requires data annotation in target domain, which is costly and labor intensive. The other approach is to generate synthetic data that is analogous to the distribution of target domain [138, 139]. Although this approach could provide unlimited synthetic training data, the model trained may not perform well as compared to real data with much more complicated distributions.

In this work, we focus on the image classification task and aim to solve the unsupervised domain adaptation problem. In our problem setting, the source domain contains a large amount of annotated data, but there is no annotation available for the images in the target domain. The two domains share the same high level categories although they are drawn from different distributions.

We propose Factorized Adversarial Networks (FAN) to address this unsupervised domain adaptation problem. FAN encodes input data from both domains to a latent

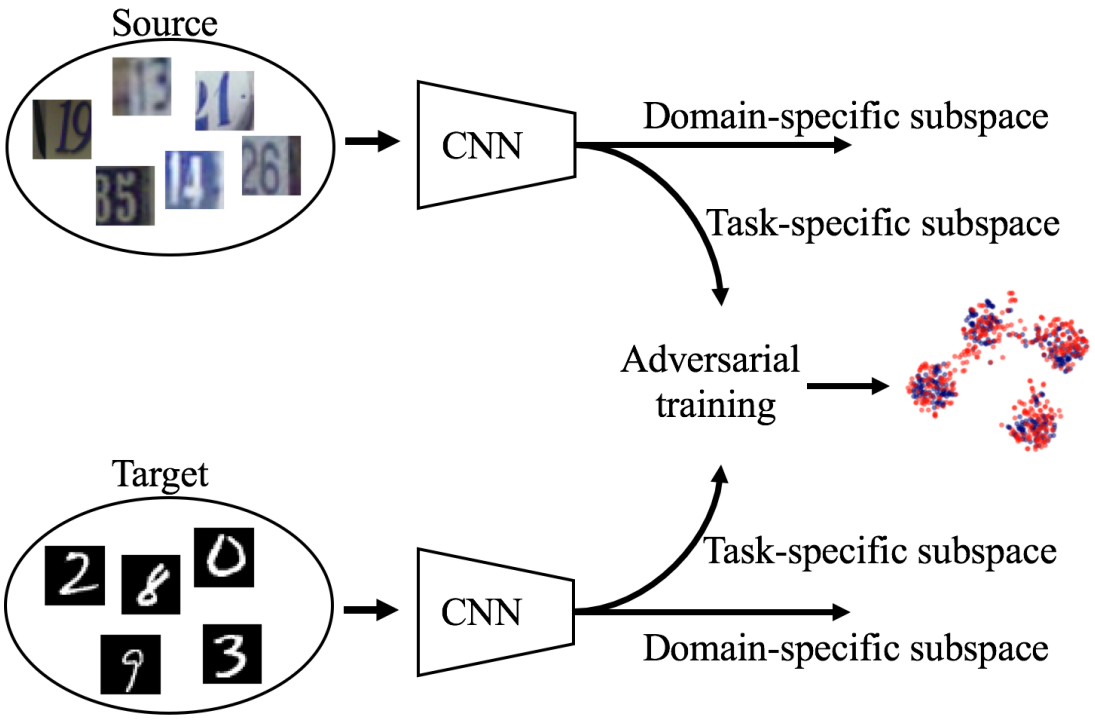


Figure 5.1: The proposed unsupervised domain adaptation approach factorizes source and target latent feature space into two subspaces using two different networks. The domain-specific subspace stores domain-specific information, while the task-specific subspace stores the category information. We use adversarial training to minimize the discrepancy between the two task-specific subspaces.

embedding space which is factorized into two complementary subspaces, a domain-specific subspace (DSS) and a task-specific subspace (TSS), as illustrated in Figure 5.1. In an image recognition scenario, the task-specific subspace should ideally only contain image category related information, while the domain-specific subspace contains domain characteristics that are irrelevant to classification, e.g., different backgrounds should not impact digit recognition. We use a mutual information loss to enforce the orthogonality constraint between the two subspaces. The motivation of this factorization is to allow us to adapt only the task-specific subspace of the target domain to that of the source domain. In order to do the adaptation, we apply an adversarial network to minimize the distribution discrepancy between the two task-specific subspaces, with loss function adopted from the Generative Adversarial Network (GAN) [45].

A two-stage training process is used to train our FAN. In the first stage, we train a convolutional network in source domain to predict the image labels as well as reconstruct the input images. The features in task-specific subspace are used to predict the image labels, while the domain-specific subspace features, concatenated with the image classification logits, are used to reconstruct the input images. In the second stage, we train the network in target domain using the adversarial loss and reconstruction loss to generate a task-specific subspace that is indistinguishable from the one generated in source domain. A discriminator network is used to judge from which domain the task-specific features are generated. The network at target domain and the discriminator network are updated by the gradients in an adversarial way so that the task-specific subspace of target domain is adapted to that subspace of source domain.

We apply our proposed method to visual domain adaptation using the benchmark digits datasets, including MNIST [140], USPS [141] and SVHN [142], and achieve superior results compared to the state-of-the-art approaches. We also apply the method to two real-world tagging datasets that we collected, one from crawling images using search engines such as Google and Flickr, and the other from photos shot by mobile phones. The two datasets share the same 100 classes with each dataset containing more than 115,000 images and we achieve significant improvement on the classification task compared with the state-of-the-arts.

In summary, our contributions are three-fold:

- A novel Factorized Adversarial Networks to tackle the unsupervised domain adaptation in an effective way.
- Detailed analysis on the design of the network architecture along with visualization of the factorized subspaces.
- New state-of-the-art domain adaptation results on digits benchmark datasets as well as newly collected larger-scale real-world tagging datasets.

5.2 Related Work

Unsupervised domain adaptation. Extensive studies on unsupervised domain adaptation have been conducted in recent years in order to effectively transfer the representative features learned in source domain to target domain. In this section, we focus on research utilizing deep neural networks as they have a better generalization ability even for the complex distributions [112, 143, 144].

One category of unsupervised domain adaptation applies the Maximum Mean Discrepancy (MMD) [145] loss as a metric to learn the domain invariant features. The MMD loss computes the distance between the embedding spaces of two domains using kernel tricks. Deep domain confusion (DDC) [146] minimizes both classification loss and MMD loss in one layer. Deep adaptation network proposed in [147] places MMD loss at multiple task-specific layers that have been embedded in a reproducing kernel Hilbert space, while other layers are shared between source and target domains. Similarly, the domain separation network (DSN) [148] maintains a shared embedding between two domains as well as the individual domain representations. Deep Reconstruction-Classification Network (DRCN) [149] shares the encoding for both source and target domains. On the contrary, the work in [150] demonstrates that it is effective to relate the weights in the form of linear transformations instead of sharing. Unlike the above discussed approaches, the authors in [151] proposed deep correlation alignment (CORAL) algorithm to match the covariance of the source features and the target features to learn a transformation from the source domain to the target domain.

Based on the idea of adversarial training [45], several studies propose using a domain classifier built on top of the networks to distinguish the represented features from the two distributions. Features extracted from the two domains are utilized to train the domain classifier, along with the classification loss for the source domain [152]. The gradient reversal algorithm (RevGrad) algorithm [153] trains the domain classifier by reversing its gradients. The authors of [33] propose an adversarial discriminative domain adaptation (ADDA) model in which weights are not shared between the source and target domains, and the network in target domain is trained to fool the domain classifier so that it cannot predict the two domains reliably.

Generative adversarial networks. GAN [45] related approaches are also used to synthesize images and perform unsupervised domain adaptation in the joint distribution space. A generator is trained to model the image distribution and generate the synthetic images while a discriminator is trained to differentiate the synthesized distribution and the real distribution. Coupled GAN (CoGAN) [154] uses two GANs on source and target domain to generate images from the two distributions. The two GANs have the same noise as input and domain adaptation is implemented by training a classifier on the input of the discriminator. The work in [138] uses images from source domain as a condition for the generator. Both the generated images and the source images are applied to train the classifier. The authors of [155] propose a learning strategy to generate cross domain images and train a task-specific classifier with the generated images and the source distributions.

Hidden factors discovery. There has been some research work on discovering the higher-order factors of variation from the latent space on the image classification and generation tasks [156, 157, 158, 159]. For example, the work at [156] utilizes the autoencoder to disentangle the various transformations from input distributions. The network is jointly trained to reconstruct input images as well as estimate the image category. On the contrary, InfoGAN [157] is proposed to learn disentangled representations from images in an unsupervised fashion by decomposing the latent code from input noise vector. In this study, we propose learning the task-specific feature in an effective way instead of learning interpretable hidden factors, and we find

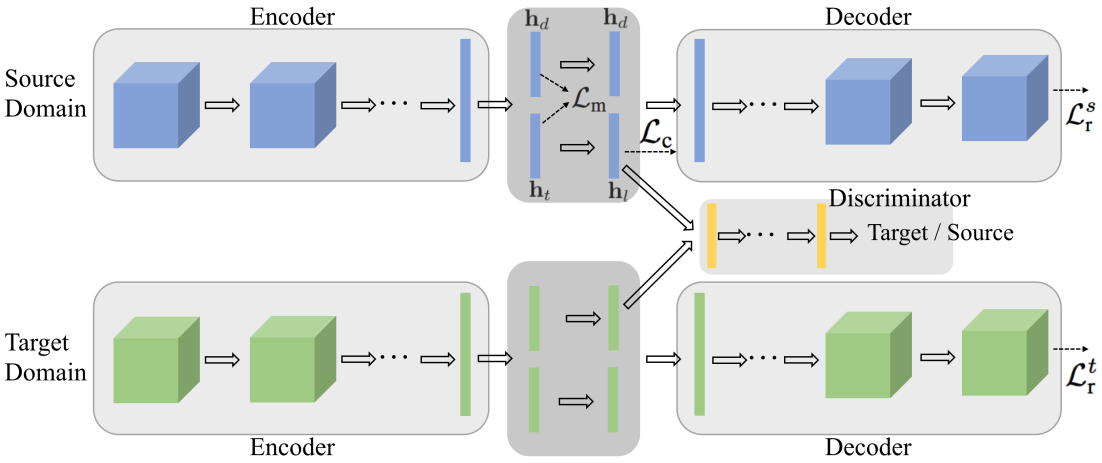


Figure 5.2: The architecture of FAN. The encoders from two domains map input images into two feature spaces. Both feature spaces are factorized into two subspaces, the domain-specific subspace (DSS) and the task-specific subspace (TSS). The adaptation is accomplished by jointly training the discriminator and target network using both the GAN loss and reconstruction loss to find the domain invariant feature in TSS.

that factorizing the domain representations helps to adapt the knowledge between two domains.

Comparison with similar studies. The motivation of our proposed FAN is to find a subspace where unsupervised domain adaptation for classification is most appropriate. It shares similarities with previous studies, especially DSN [148] and the ADDA [33]. While domain separation [156, 160, 161] and adversarial training [138, 152] have been extensively explored in many tasks in existing liturature, we unify the two appoaches in one novel framework for unsupervised domain adaptation and demonstrate its clear advantage over DSN [148] and ADDA [33] in experiments.

5.3 Our Approach

In this section, we present our Factorized Adversarial Networks (FAN) for unsupervised domain adaptation. The architecture of FAN is illustrated in Figure 5.2, where we have two encoder-decoder structured neural networks, one for source domain and one for target domain, that mirror each other except for the training losses, as well as a discriminator network. We aim to find a domain invariant feature space that retains

the classification information through adversarial training. To achieve this, we explicitly factor the latent feature space into a task-specific subspace and a complementary domain-specific subspace, where the task-specific subspace aims to minimize the classification loss across domains while the domain-specific subspace combined with classification logits targets at reconstructing the input samples. The task-specific subspace, if indistinguishable by the adversarial discriminator which domain it comes from, should retain the classification information invariant to domain shifts; the domain-specific subspace, on the other hand, should capture the domain-specific but classification-irrelevant information for reconstruction. The proposed explicit feature space factorization helps to remove some domain-specific information and relieve the burden of adversarial training for more effective domain adaptation.

More formally, in our unsupervised domain adaptation, we have a source distribution $\mathcal{S}$ that includes N^s labeled images $\{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1}^{N^s}$ where $\mathbf{y}_i^s$ is a one-hot vector encoding the image class label, and a target distribution $\mathcal{T}$ contains N^t unlabeled images $\left\{(\mathbf{x}_i^t)\right\}_{i=1}^{N^t}$. Our goal is to first find a mapping M^s that maps the source task-specific subspace to the source logit space with labeled training data, and then find a mapping function M^t for the target domain that maps the target task-specific subspace into the target logit space that is indistinguishable from the source logit space. The target mapping function M^t thus retains the discriminative information needed for target domain, and therefore, inference in target domain could be easily done with M^t and *softmax*. Our learning procedure consists of two steps: we first train a source domain network that factors the latent feature space, and we then update the target domain network by adapting the target domain task-specific subspace to its source domain counterpart with the help of adversarial training. We discuss these two steps in the following sections.

5.3.1 Feature Space Factorization

Our networks contain two convolutional encoder-decoder networks and the latent feature space generated by the encoders is factorized into complementary task-specific subspace and domain-specific subspace. In the first step of our approach, we train

our factorization network in source domain as shown in Figure 5.2. To avoid cluttered notations, we drop domain indicator superscripts s in the following when there is no confusion. Let $\mathbf{h} = Enc(\mathbf{x}; \theta_e)$ denote the encoder function Enc that encodes the input sample $\mathbf{x}$ into a latent feature $\mathbf{h}$ with parameter θ_e in source domain. We split the latent feature $\mathbf{h}$ into two parts $\mathbf{h}_d$ and $\mathbf{h}_t$, where $\mathbf{h}_d$ represents the feature in the domain-specific subspace and $\mathbf{h}_t$ represents the feature in the task-specific subspace. The mapping $\mathbf{h}_l = M(\mathbf{h}_t; \theta_m)$ maps the task-specific subspace into a logit space with parameters θ_m. We then concatenate $\mathbf{h}_d$ and $\mathbf{h}_l$ and feed them into a decoder $Dec(\mathbf{h}_d, \mathbf{h}_l; \theta_d)$ to reconstruct the input sample $\mathbf{x}$, where $\mathbf{h}_l$ includes the necessary attributes for reconstruction. Ideally, $\mathbf{h}_t$ should contain discriminant information that is invariant to different domains while $\mathbf{h}_d$ retains information that is specific to the domain, less relevant to classification but necessary for reconstruction. We optimize the following objective function in order to obtain the two desired subspaces in source domain:

$$\mathcal{L}_{\text{source}} = \alpha \mathcal{L}_{\text{c}} + \beta \mathcal{L}_{\text{m}} + \mathcal{L}_{\text{r}} \tag{5.1}$$

where α, β are hyper parameters that control the trade-off among loss terms.

$\mathcal{L}_{\text{c}}$ is the cross-entropy loss to train the source network for classification with the parameters $\{\theta_e, \theta_m\}$ using source domain labeled training data.

$$\mathcal{L}_{\text{c}} = -\sum_{i=1}^{N} \mathbf{y}_i \cdot \log \hat{\mathbf{y}}_i \tag{5.2}$$

where $\hat{\mathbf{y}}$ is the *softmax* output of the classification branch, $\hat{\mathbf{y}} = softmax(M(\mathbf{h}_t; \theta_m))$.

We add a mutual information loss term $\mathcal{L}_{\text{m}}$ to encourage orthogonality between the domain-specific subspace and task-specific subspace:

$$\mathcal{L}_{\text{m}} = \sum_{i=1}^{N} \left\| \mathbf{h}_{ti}^{\mathbf{T}} \mathbf{h}_{di} \right\|^2 \tag{5.3}$$

where $\mathbf{h}_{ti}$ and $\mathbf{h}_{di}$ denote the domain-specific feature and task-specific feature for the i-th sample, respectively.

We use the reconstruction loss $\mathcal{L}_{\text{r}}$ to minimizes the squared error between the input sample and the reconstructed one:

$$\mathcal{L}_{\text{r}} = \sum_{i=1}^{N} \left\| \mathbf{x}_i - Dec(\mathbf{h}_{di}, \mathbf{h}_{li}; \theta_d) \right\|^2 \tag{5.4}$$

where $\mathbf{h}_{li}$ denote the logit vector for the i-th sample.

The three loss terms play together in the optimization of Eqn. 5.1. The classification loss $\mathcal{L}_{\mathrm{c}}$ encourages the learned feature $\mathbf{h}_t$ to retain discriminative information as much as possible, the reconstruction loss $\mathcal{L}_{\mathrm{r}}$ relies on domain-specific information from $\mathbf{h}_d$ with the logit input $\mathbf{h}_l$ for reconstruction, and the mutual loss $\mathcal{L}_{\mathrm{m}}$ encourages the separation of the two subspaces. Thus we can obtain a task-specific space $\mathbf{h}_t$ that is discriminative with much less domain-specific information, and hence more invariant to domain shifts.

Without duplicate elaboration, the target domain network holds the same architecture as the source domain network. In the second step of our approach, we fix the learned source domain factorization network and train the target factorization network with adversarial adaptation, as discussed in following section.

5.3.2 Adversarial Domain Adaptation

Our factorization network is designed to capture discriminant information in the task-specific subspace while dropping domain-specific information as much as possible. We leverage adversarial training to minimize the discrepancy between the task-specific subspace of the target domain and that of the source domain so that we can easily transfer the knowledge learned from source domain to target domain. Specifically, we learn our target domain neural network by optimizing the following objective function:

$$\mathcal{L}_{\text{target}} = \mu\mathcal{L}_{\text{adv}_{\mathbf{D}}} + \nu\mathcal{L}_{\text{adv}_M} + \mathcal{L}_{\mathrm{r}} \tag{5.5}$$

where μ and ν are the hyper parameters that balance the contributions of adversarial training loss.

The reconstruction loss $\mathcal{L}_{\mathrm{r}}$ in target domain is similarly defined as Eqn. 5.4 over target domain network parameters. The adversarial training losses are defined similarly to the GAN loss [45]. Instead of using the task-specific subspace directly, we use the logit space obtained from the source domain to guide the learning in the target domain, which works better in practice. The discriminator $\mathbf{D}$ maps the input logit space into a binary label, where "true" denotes the source domain and "false" denotes the target domain. The target domain network is learned in an adversarial way to fool the discriminator

so that the discrepancy between the two logit spaces is minimized. Specifically, the adversarial losses $\mathcal{L}_{\text{adv}_\mathbf{D}}$ for optimizing the discriminator $\mathbf{D}$ and $\mathcal{L}_{\text{adv}_M}$ for optimizing the target domain encoder are defined as

$$\min_{\mathbf{D}} \mathcal{L}_{\text{adv}_\mathbf{D}} = -\mathbb{E}_{\mathbf{x}_s \sim \mathcal{S}} \log \mathbf{D}(M^s(\mathbf{h}_t^s; \theta_m^s)) - \mathbb{E}_{\mathbf{x}_t \sim \mathcal{T}} \log(1 - \mathbf{D}(M^t(\mathbf{h}_t^t; \theta_m^t)) \tag{5.6}$$

$$\min_{\Theta} \mathcal{L}_{\text{adv}_M} = -\mathbb{E}_{\mathbf{x}_t \sim \mathcal{T}} \log(\mathbf{D}(M^t(\mathbf{h}_t^t; \theta_m^t))) \tag{5.7}$$

where Θ denote the network parameters for the target domain encoder and logit mapping. As the task-specific subspace at target domain aims to learn a similar distribution as the one from source domain, the mutual information loss is not necessary for the target domain. In the experiments, we did try using Eqn. 5.3 at target domain , but did not observe further improvement.

Unlike the symmetric structure of our network as demonstrated in Figure 5.2, we perform asymmetric adaptation during optimization where the target domain network is fine-tuned from source domain network instead of weight sharing for the two networks. Previous efforts explored using shared weights between source and target networks to reduce model parameters [34, 152], or leave the target network completely untied [150, 33]. We found that it is not necessary to share the weights for shallow networks such as LeNet [140], but imperative to partially share some early network layers for deeper neural networks, such as ResNet [143], which is the standard practice to train the deep nets. By jointly optimizing the adversarial loss and reconstruction loss, we force the target domain task-specific subspace to match the distribution of the source domain task-specific subspace, which is discriminative for the classification task, while leaving the less relevant target domain-specific representations for the domain-specific subspace to capture. Together, the two terms encourage the network to learn more discriminative and domain invariant feature representations for the task.

5.4 Experiments

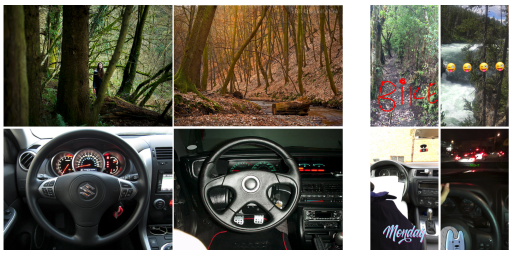
We evaluate the proposed FAN on the tasks of unsupervised domain adaptation using benchmark datasets including MNIST [140], USPS [141] and SVHN [142], as well as much larger real-world tagging datasets we collected that contain more than 100,000 images, respectively. We demonstrate that our approach is significantly improved compared to previous state-of-the-art methods.

5.4.1 Digits Datasets

We use three digits datasets, MNIST [140], USPS [141] and SVHN [142], as the benchmark and follow the previous studies [149, 152, 153, 33, 154] to perform three unsupervised adaptation settings including MNIST $\rightarrow$ USPS, USPS $\rightarrow$ MNIST and SVHN $\rightarrow$ MNIST. The benchmark datasets contain images of 10 digits ranging from 0 to 9. Some sample images from the three datasets are shown in Figure 5.3a. To run experiments in an unsupervised manner, the labels of the target domain training images are withheld.

Network architecture The network we use in the experiments contains an encoder and a decoder and has the same structure under the three experiment settings. Following the recent work [33] for fair comparison, we adopt a similarly modified LeNet [140] as the encoder that differs only in utilizing batch normalization (BN). We also applied BN for [33] but observed no improvement. Specifically, the encoder consists of two convolutional layers with kernel size 5 and the number of filters 20 and 50, respectively. Each convolutional layer is followed by rectified linear units (ReLU), BN, and max pooling layers. After that we have two fully connected (FC) layers with 500 and 100 hidden units respectively. The activations from the last FC layer is split into two parts, one for domain-specific subspace and the other for task-specific subspace. The task-specific feature is connected to an FC layer to get the classification logits for prediction, while the domain-specific feature is concatenated with the classification logits as input for decoding phase. The decoder employs a deconvolution architecture [162] including one FC layer with 300 hidden units, two $5\times5\times16$ convolutional layers, one upsampling layer to 28×28, and two 3×3 convolutional layers with 16 and 1 filters, respectively.

(a) Example images from the three digits datasets. Left three columns: MNIST; middle three columns: USPS; right three columns: SVHN.



(b) Example images from the *Crawling* dataset (left two columns) and the *Mobile* dataset (right two columns). Top row: forest; bottom row: steering wheel.

Figure 5.3: Visualization of example images from the five datasets used in the study.

The FC layers and convolutional layers are followed by ReLU and BN, except for the last convolutional layer that gives the reconstruction output. The logit activations from the two domains are sent to the discriminator network which contains three FC layers. The first two FC layers have 500 hidden units followed by ReLU and BN. The last FC layer provides the domain label estimation for the input samples.

Implementation details Since images in different datasets varies in size, we resize the images in USPS and SVHN datasets to 28×28 in order to match the input image size in MNIST. In addition, we convert the RGB images from SVHN to gray scale images. All the pixel values are normalized to a range of 0 to 1. For the unsupervised adaptation between MNIST and USPS, two training paradigms are implemented. The first one follows the training strategy introduced in [163], which sampled 2,000 training images from MNIST and 1,800 training images from USPS. For the second training protocol, we consider utilizing all the training data from the two domains and denote it as MNIST→USPS (full) and USPS→MNIST (full). For both training protocols, the testing set remains the same. For adaptation from SVHN to MNIST, we use all the training images from the two datasets. The training process contains two steps. The first step is to train a model in the source domain using Eqn. 5.1 with α as 2 and β as 1. In the second step, we fix the trained model in source domain and train the recognition model in the target domain using the Eqn. 5.5, where μ is 2 and ν is 1. We initialize the target domain network using the weights of the model trained in source domain. No data augmentation setting is utilized in the experiments.

Table 5.1: Experimental results on unsupervised domain adaptation for the digits datasets including MNIST, USPS, and SVHN. Full denotes using the entire training set for the domain adaptation between MNIST and USPS. The last column shows the largest improvement over each method out of the three experiments.

Method	MNIST→USPS	USPS→MNIST	SVHN→MNIST	Largest Improvement
Baseline	0.752 ± 0.016	0.571 ± 0.017	0.601 ± 0.011	0.339
DSN[148][164]	0.913	-	0.827	0.098
RevGrad[153]	0.771 ± 0.018	0.730 ± 0.020	0.739	0.186
DDC[152]	0.791 ± 0.005	0.665 ± 0.033	0.681 ± 0.003	0.245
CoGAN[154]	0.912 ± 0.008	0.891 ± 0.008	-	0.019
DRCN[149]	0.918 ± 0.0009	0.737 ± 0.0004	0.820 ± 0.0016	0.173
ADDA[33]	0.894 ± 0.002	0.901 ± 0.008	0.760 ± 0.018	0.165
Ours	**0.921 ± 0.014**	**0.910 ± 0.011**	**0.925 ± 0.011**	-
Ours (full)	**0.963 ± 0.002**	**0.971 ± 0.008**	-	-

Comparison results Table 5.1 shows our results as compared with recent methods. Our approach clearly achives the best overall performance on all three domain adaptation experiments under the same settings. Compared with previous methods, our method significantly outperforms each of them at least on one of the three experiments, with a gap of over 10% in many cases, as shown in the last column in Table 5.1. For the adaptation between MNIST and USPS, we also show results using the full set of training data from both domains and observe that it significantly improves the accuracy, implying that our adaptation network can better minize the distribution shift with more training data.

Ablation analysis of our network design We conduct ablation study on the design of our factorization architecture. The structure for four network settings are shown in Figure 5.4 with the following details.

- **Joint feature**: As shown in Figure 5.4a, we learn a joint feature space for both image reconstruction and classification, and use reconstruction losses in both domains along with the classification loss in source domain to train the network.
- **Feature separation**: As shown in Figure 5.4b, in this setting, we separate the latent features into two parts. One part is used for reconstruction and the other

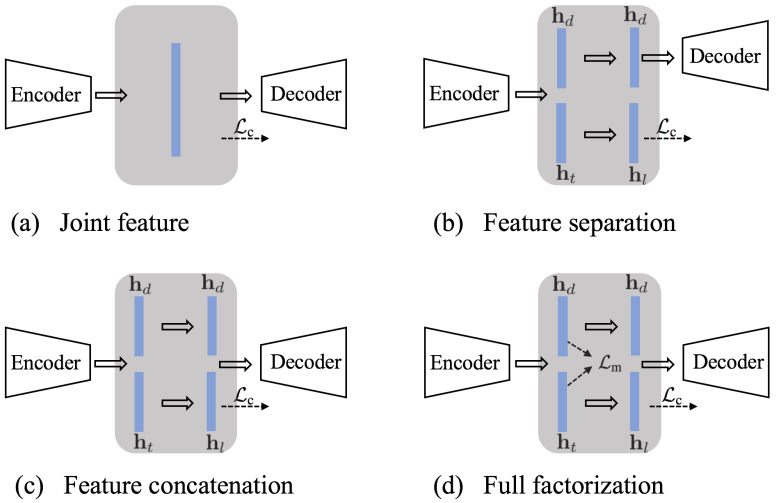


Figure 5.4: Four network architectures for the study of feature factorization.

part is used for classification.

- **Feature concatenation**: As shown in Figure 5.4c, the previous reconstruction features are concatenated with the classification logits as new reconstruction features.

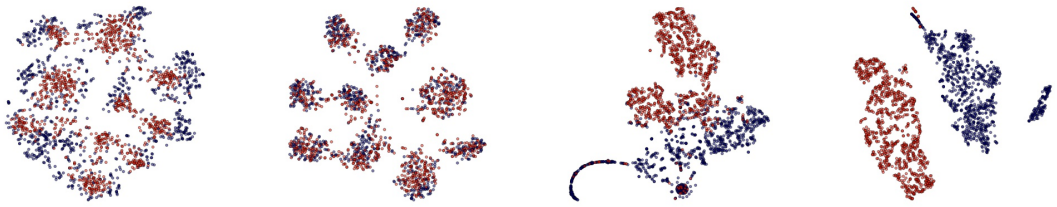
- **Full factorization**: As shown in Figure 5.4d, we add mutual information loss in this setting to explicitly enforce the orthogonality between the two separated features, thus factorizing the latent feature space into a domain-specific subspace and a task-specific subspace.

For all four settings, we conduct the same two-stage training process and apply the adversarial learning at the second stage. The results shown in Table 5.2 indicate that we could obtain stronger results by better separating the features, and our factorization method yields the best results.

Analysis of the embedding spaces Besides the quantitative results, we visualize the high-dimensional features of the factorized subspaces in the 2D plane for adaptation from SVHN to MNIST using the t-SNE [165]. We randomly select 1,000 images from the two testing sets and show visualization results in Figure 5.5. We set perplexity to 35 for all four visualization results. The embedding of the logits space before and after

Table 5.2: Analysis of the effects of feature factorization under different network structures.

Method	Joint feature	Feature separation	Feature concatenation	Full factorization
MNIST → USPS (full)	0.955 ± 0.004	0.958 ± 0.002	0.961 ± 0.002	**0.963 ± 0.002**
USPS → MNIST (full)	0.933 ± 0.017	0.936 ± 0.014	0.958 ± 0.009	**0.971 ± 0.008**
SVHN → MNIST	0.829 ± 0.019	0.858 ± 0.024	0.905 ± 0.006	**0.925 ± 0.011**



(a) Embedding of the logit space before the adaptation.

(b) Embedding of the logit space after the adaptation.

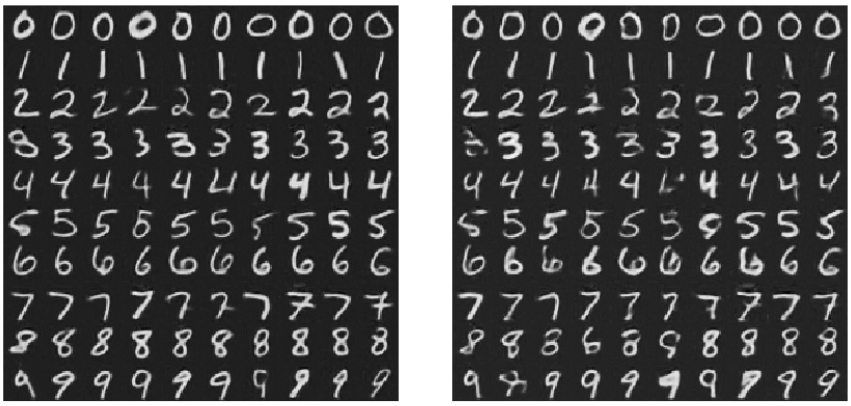
(c) Embedding of the domain-specific subspace before the adaptation.

(d) Embedding of the domain-specific subspace after the adaptation.

Figure 5.5: Visualization of the domain adaptation from SVHN (source domain, red color) to MNIST (target domain, blue color). We show the visualization of t-SNE embedding for the logits space before adaptation (a) and after adaptation (b), and the domain-specific subspace before adaptation (c) and after adaptation (d).

adaptation for the two domains are shown in Figure 5.5a and Figure 5.5b, respectively. As expected, after adaptation, the samples from the target domain are clustered into more obvious groups and match better with the clusters in the source domain.

The visualization of the domain-specific subspaces before and after adaptation are shown in Figure 5.5c and Figure 5.5d, respectively. After adaptation, we simultaneously learn a good task-specific subspace on the target domain and a good domain-specific subspace. The domain-specific subspace should capture information specific to the domain, and therefore, the two domain-specific subspaces are further divided after adaptation, which proves our learning algorithm is effective.



(a) (b)

Figure 5.6: Reconstruction results using the target domain reconstruction network for domain adaptation from SVHN to MNIST. (a) Reconstruction results using the testing samples from target domain. (b) Reconstruction results using the concatenation of domain specific features from target domain and classification logits from source domain.

Furthermore, we analyze the embedding subspaces of the target domain by showing two reconstruction results. Figure 5.6a demonstrates the reconstruction results using features extracted from target domain testing samples. We also concatenate the domain-specific features of the target domain samples with the logits activations of randomly selected testing images from source domain of the same class, and show the reconstruction results in Figure 5.6b. Although reconstruction quality of Figure 5.6b is not as good as that of Figure 5.6a, the images are still very similar, which proves that task-specific subspaces for the two domains indeed share similar distributions and that the target domain-specific subspace stores the domain characteristics for reconstruction.

5.4.2 Real-world tagging Datasets

While many studies in the literature tackle unsupervised domain adaptation, they mostly evaluate their algorithms on small and simple datasets such as digits dataset [140, 141, 142], and the office dataset [166]. The capasity for domain adaptation algorithms to work for large-scale real-world complex applications remains unclear. Previous work [148] points out some problems with evaluation on office dataset [166, 167], where pre-trained models from ImageNet have to be used [168]. So instead of working on the toy

office dataset, we collected two real-world tagging datasets to benchmark unsupervised domain adaptation algorithms, where we have sufficient images to train deep networks from scratch.

The first dataset is collected from the search engines and named *Crawling* dataset, while the second dataset is collected from the photos shot by mobile phones and titled *Mobile* dataset. The two datasets contain the same 100 classes. Some example images from the two datasets are shown in Figure 5.3b. There are two major differences between the two datasets: 1) the images in *Crawling* dataset usually have good quality and clear background while the images in the *Mobile* dataset suffer from several defects such as image blur and out of focus, as well as noisy background and various image filters and stickers; 2) the *Mobile* dataset contains mostly vertical images while the images from the *Crawling* dataset have various image ratios. We use the *Crawling* data as the source domain and the *Mobile* data as the target domain because we can easily collect crawling data with labels by keyword searching. The *Crawling* dataset includes 150,000 training images. The *Mobile* dataset contains 115,000 images out of which we randomly select 100,000 images as the training set, 10,000 images as the testing set, and others as the validation set. Compared with the digits datasets, the real-world tagging datasets not only have a larger scale but also are more suitable for the study on real-world scenarios.

Network architecture The encoder part of our network uses the ResNet-50 [143] architecture. The activations from the last average pooling layer are factorized equally into two parts. The task-specific subspace features are followed by a FC layer to estimate the classification logits, and the domain-specific features are concatenated with the classification logits to serve as the input for the decoder. The decoder network uses architecture from DCGAN [169]. It contains 5 fractionally-strided convolutions layers with 256, 256, 128, 64 and 3 filters respectively. Each layer is followed by ReLU and BN, except for the last layer. The discriminator network contains three FC layers. The first two FC layers have 1024 and 2048 hidden units respectively, followed by ReLU and BN. The last FC layer output is used for label domain classification.

Implementation Detail All images are resized to 256×256 and randomly cropped

Table 5.3: Top-1 and Top-5 accuracies on the testing set of the *Mobile* dataset.

Method	Top-1	Top-5
No adaptation	0.3571	0.6607
ADDA[33] (full set of target training)	0.4386	0.7533
Ours (10% of target training)	0.3946	0.6976
Ours (50% of target training)	0.4041	0.7018
Ours (full set of target training)	**0.4632**	**0.7838**

to 224×224 during the training process. α is set to 5 and β is set to 1 for Equation 5.1. And we set μ as 2 and ν as 1 for Eqn. 5.5.

In order to measure whether more unlabeled training images in target domain would contribute to the generalizability of the target model, we perform three sets of experiments in addition to that without adaptation. In the first two sets, we randomly select 10% and 50% images from each class of the target training set, while in the third set, we use the full target training set. The Top-1 and Top-5 accuracy for the testing set of the target domain are shown in Table 5.3. Compared with the model without adaptation, using 10% training images from each class could improve the the Top-1 and Top-5 accuracy as 3.75% and 3.69% respectively. Using the full training set improves the Top-1 accuracy by more than 10% and Top-5 accuracy more than 12%. We also compared our results with ADDA [33] using the full target training set, as shown in Table 5.3. Our approach outperforms ADDA on both Top-1 and Top-5 accuracy as 2.46% and 3.05% respectively. These results demonstrate that our method can significantly improve the performance over baselines in real-world applications. In addition, we show that more unlabeled training data from the target domain helps the unsupervised adaptation.

5.5 Conclusion

In this work, we introduce FAN for unsupervised domain adaptation. We factorize the latent feature space into task-specific subspace and domain-specific subspace for both source and target domains and consider the domain adaptation only on task-specific subspace. The network in source domain is jointly trained with image classification

and reconstruction under the factorization architecture to learn the discriminative task-specific subspace while pushing away domain-specific information as much as possible. The network in target domain is learned under the same factorization structure with GAN loss to adapt the target domain task-specific subspace to the source domain task-specific subspace. We evaluate our proposed framework on four domain adaptation tasks, all achieving state-of-the-art results. For future work, we would like to extend our algorithm to other vision tasks beyond image classification.

Chapter 6

Unsupervised Domain Adaptation forClassification of Histopathology Whole-Slide Images

6.1 Introduction

Advances in whole-slide scanner technology have increased the speed and reliability with which histopathology slides and other microscopic specimens are digitized. As a result of these improvements, there has been a sharp increase in the number of investigators and health-care providers adopting the use of these devices in routine research and clinical workflows. The sheer volume of digitized specimens now being generated at both small and large institutions has grown accordingly. Once digitized, these specimens are well suited for the application of sophisticated pattern recognition and machine-learning algorithms and strategies that can facilitate automated decision-support and computer-assisted diagnosis. Over the course of hundreds of years, scientists and pathologists have gone to great length to develop and optimize staining methods that augment and enhance the contrast of biological components of interest within these samples at the tissue, cell and sub-cellular levels. Hematoxylin & Eosin (H&E) is a popular stain that is applied to specimens, routinely, that results in nuclei exhibiting a bluish color with cytoplasmic regions rendered in pink [170]. In spite of the best efforts of the technicians preparing the specimens, however, slight variations in the manner in which these stains are applied to specimens often results in histopathology sections that are inconsistent in visual appearance and samples often containing processing artifacts. While there have been many attempts to completely standardize these methods, the current technology still grapples with these challenges [171, 172, 133, 173]. Since these inherent issues described can lead to variations in the results obtained using image-based quantification approaches to analyze the specimens, our team has been investigating new

methods to remove color variation across digitized specimens originating from different institutions as well as batches of imaged specimens that may have been acquired at a single institution at different time points. In earlier attempts to mitigate the color normalization issue, some investigators chose to convert the color images into gray-scale versions before performing quantitative analysis [49, 174, 175, 176, 177, 174]. However, the conversion from color space to grayscale eliminates some informational content from the digitized specimens that may be essential for rendering proper classifications and accurate diagnosis.

While the noted color variations in digital specimens present formidable technical challenges for any image analysis algorithm, mechanical distortions that can sometimes be introduced during tissue sectioning and slight variations in the underlying morphologic and structural patterns within imaged specimens can further complicate the process of automating classifications [48, 49, 178, 25, 26, 179]. In spite of all of the difficulties, investigators throughout the scientific community continue to pursue this line of research because of the the potential impact that automated, computer-aided analyses could have in clinical practice and investigative research by accelerating the throughput while reducing or eliminating the negative effect of inter- and intra-observer variations during the assessment of microscopic images. Methods based on convolutional neural networks (CNN) are currently considered state-of-the-art due to the high performance rates recently reported by some recent investigations [78, 180, 181]. Most of these studies, however, focused on supervised classification. Unfortunately, supervised classification models used on one annotated dataset (source domain) may render ineffective for another set (target domain) collected at a different institute. A widely used approach to address the challenge is to label new images on the target domain and fine-tune the model trained on source domain [32]. In fact, methods that can learn from existing datasets and adapt to new target domains, without the need for additional labeling, are among the most desirable approaches because they lend themselves to high-throughput clinical environments and big data research experiments involving large patient cohorts [46].

In this study, we aim to address the challenges presented by variations in staining,

morphologic and architectural profiles within histopathology whole-slide images (WSIs) in a completely unsupervised manner. We use two approaches to achieve knowledge transfer from the source domain to the target domain. In the first approach, we adopt two off-the-shelf color normalization [55, 182] on the images from the target domain, where the model learned from the source domain is applied to the target images after being normalized to the reference image chosen from the source domain. In the second approach, we adopt an unsupervised domain adaptation paradigm to align the image distributions along the annotated source domain and the unlabeled target domain [33, 34]. We apply adversarial training to minimize the distribution discrepancy in the feature space between the domains, using the loss function adopted from the Generative Adversarial Network [45]. We subsequently develop a Siamese architecture for the target network to serve as a regularization of patches within the WSI's. We validate the proposed methods on a set of publicly available histopathology datasets and then further test performance using a new dataset that is collected locally at Rutgers Cancer Institute of New Jersey. The experimental results show the merit of these strategies.

6.2 Related Works

6.2.1 Color Normalization

In an attempt to address the challenge of the previously described color batch effects, many investigators have applied color normalization methods to the imaged histopathology specimens prior to analysis [183, 184, 185, 133, 182, 186, 187, 188, 189, 190, 191, 192]. One common approach for analyzing tissue samples is to treat stains as agents exhibiting selective affinities for specific biological substances. With an implicit assumption that the proportion of pixels associated with each stain is same in source and target images, histogram-based methods are investigated [193, 194, 195, 48, 196, 197, 198, 186, 199]. The main drawback of histogram-based methods is that they often introduce visual artifacts into the resulting images. Color deconvolution strategies [55, 200, 201] have been utilized extensively in the analysis imaged histopathology specimens by separating RGB images into individual channels such as by converting

from RBG to Lab [171] or HSV space [202]. The limitation of this approach is that both the image-specific stain matrix and a control tissue stained with a single stain is required to perform the color deconvolution. Another strategy that has been explored is to utilize blind color decomposition which is achieved by applying expectation and maximization operations on color distributions within the Maxwell color triangle [201]. This strategy requires a heuristic randomization function to select stable colors for performing the estimation, thus it is prone to be affected by achromatic pixels at the weak stain pixels. Tissue inherent morphological and structural features may not be preserved after color deconvolution since statistical characteristics of decomposition channels are modified during this process. Model-based color normalization has also been studied in such applications by including Gaussian mixture models [203, 198, 171, 133, 173], matrix factorization [182], sparse encoder [190], and wavelet transformation with independent component analysis [187]. Other studies utilize generative models [45] to achieve the stain normalization [204, 205, 206, 207]. Typically, a reference image is needed from a group of image dataset. The different reference image would give the different domain adaptation performance. Color normalization models can provide stain estimation, but they are solely dependent on image color information, while the morphology and spatial structural dependency among imaged tissues is not considered [201, 199, 202, 186], which could lead to unpredictable results especially when strong staining variations appear in the imaged specimens.

6.2.2 Adversarial Domain Adaptation

In recent years, there have been many studies on unsupervised domain adaptation for transferring the learned representative features from the source to the target domain [208, 209, 164, 210]. The works based on CNN show significant advantages due to better generalization across different distributions [112, 144]. With the development of the Generative Adversarial Networks (GAN) [45], studies show the synthesized images could be used to perform unsupervised domain adaptation in a learned feature space where a generator is applied to learn the image distribution and generate the synthetic images while a discriminator is trained to differentiate the synthesized and the real

distribution [154, 138]. For example, Generate-to-Adapt [189] proposes to learn a joint embedding space between the source and target domain, where the embedding space could be used to synthesize both the source and target images. Inspired by previous studies, we utilize the adversarial training to find a discriminative feature space that can be used to transfer the knowledge from source to target domain. Furthermore, we introduce a Siamese architecture at target domain which can be used to regularize the classification of WSIs in an unsupervised manner.

6.3 MATERIALS

For the purposes of the current study, we focus on unsupervised domain adaptation of imaged prostate cancer histopathology specimens. Prostate cancer is the most common non-cutaneous malignancy afflicting 1 in 7 men in the United States [211]. Over the years, Gleason scores have consistently served as a reliable predictor for differential prostate cancer diagnosis [178]. Unfortunately, Gleason grading can be extremely time-consuming when attempting to systematically evaluate large, giga-pixel-sized WSIs. Furthermore, inter- and intra-observer variability errors often arise when pathologists are called upon to render diagnoses based on WSIs. In order to provide an objective and reproducible Gleason grading score on such datasets, reliable computational methods are required for detection, extraction, and recognition of the underlying histopathological patterns. Much of the progress in this area of research has focused on supervised classification of the imaged tissues [48, 102, 212, 51, 53]. However, the fact that histopathology WSIs obtained from different institutions often present divergent glandular appearances due to the fact that the acquisitional and optical properties of the specific type of scanners used and differences in the sectioning and staining procedures utilized introduce significant variations in the resulting images. Additionally, WSIs scanned by from different institution may have different image resolution as they were scanned under various microscopy. Figure 6.1 shows representative prostate cancer tissue images originating from different institutions. Note the variations in glandular distributions and staining appearance.

Our team investigated the use of unsupervised domain adaptation for histopathology

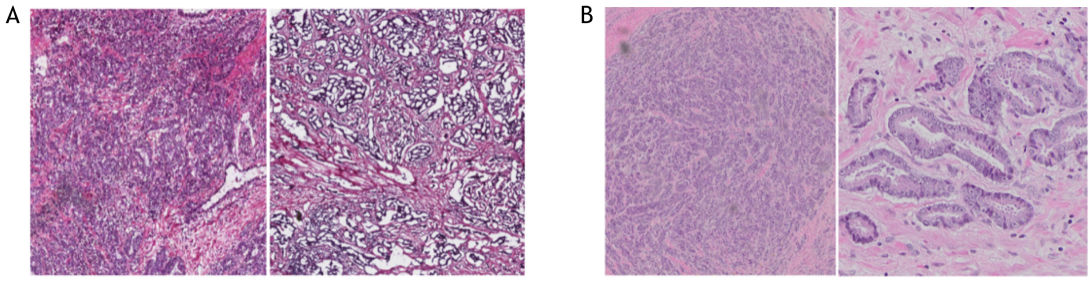


Figure 6.1: Examples of prostate cancer histopathology WSIs from TCGA (A) and RCINJ (B). The WSIs from different institutes present different glandular distribution and staining appearance.

images and tested the approach on two datasets. The first which is publicly available is called The Cancer Genome Atlas (TCGA) dataset [67]. The other is a dataset collected locally at Rutgers Cancer Institute of New Jersey (RCINJ) after obtaining institutional review board (IRB) approval. All the histopathology images are H&E stained. For the first setting of unsupervised domain adaptation, we only use the TCGA dataset. The TCGA prostate cancer dataset includes histopathology WSIs uploaded from 32 institutions that have been acquired at 40× and 20× magnifications. We crop the WSIs into patches of size 2048×2048. We calculate the tissue area on the grayscale images and remove the images with tissue area less than half of the patch size. The dataset includes Gleason scores, ranging from 6 to 10, that have been annotated by pathologists. As the University of Pittsburgh (UP) had contributed more images than other institutions, we treat the UP images as the target domain where the annotations are withheld and the images from other institutions as the source domain, which we denote as TCGA (w/o UP). We show the total number of WSIs and the cropped patches from TCGA in Table 6.1 and UP in parentheses. We denote the adaptation setting as TCGA (w/o UP) → UP. For the second setting of the unsupervised domain adaptation, we use all the images from TCGA as the source domain, and images from RCINJ as the target domain. The images from RCINJ are acquired at 20× magnification. More details of the RCINJ dataset are shown in Table 6.2. The dataset was labeled as Gleason scores as 6 or 8 by a board-certified pathologist. We denote this adaptation as TCGA → RCINJ.

Table 6.1: The number of WSIs and patches of the prostate histopathology images from TCGA under different Gleason scores. The images from University of Pittsburgh (UP) are shown in parentheses.

	Gleason 6	Gleason 7	Gleason 8	Gleason 9	Gleason 10
# WSIs	115 (32)	395 (95)	94 (20)	128 (24)	4 (0)
# Patches	16293 (6517)	67162 (26583)	16204 (4968)	23978 (9606)	342 (0)

Table 6.2: The number of WSIs and patches of the prostate histopathology images from RCINJ under different Gleason scores.

	Gleason 6	Gleason 8
# WSIs	57	26
# Patches	3933	666

For the two sets of unsupervised adaptation, we aimed to transfer the knowledge gained from the source image data to the images in target domain so that a network could reliably classify the WSIs in the target domain into low- and high-Gleason score categories. Specifically, the methods were used to divide the TCGA dataset into low Gleason grade for the WSIs with score as 6 and 7, and high Gleason grade for the WSIs with score as 8, 9 and 10. In the case of the RCINJ dataset, the WSIs with Gleason score of 6 belong to the low-Gleason grade whereas those assigned a Gleason score of 8 belonging to high Gleason grade.

6.4 Methods

In this section, we introduce the two different unsupervised methods to solve the domain variation necessary for rendering accurate classification of histopathology images.

6.4.1 Problem Formulation

For the purposes of the experimental design, the annotated images are established at source domain whereas the unlabeled images are housed at the target domain. To facilitate the study, for the source domain, we denote $\mathcal{S}$ as the image distribution, N_s as the total number of annotated images, $\{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1}^{N_s}$ as the i^{th} image $\mathbf{x}^s$ with the one-hot category information of $\mathbf{y}^s$. Similarly, for the target domain, we denote $\mathcal{T}$ as

the image distribution, N_t as the total number of unlabeled images, $\{(\mathbf{x}_i^t)\}_{i=1}^{N_t}$ as the i^{th} image unlabeled image $\mathbf{x}^t$.

We use the images from the source domain to learn a mapping function M_s that can reliably transform the images to the feature space. Then we apply two approaches for the unsupervised domain adaptation. The first transfers the staining information from the images of the source domain to the images of the target domain so that the classification of target domain can be easily achieved by using M_s. The second identifies the mapping M_t that must occur at the target domain to obtain a similar feature space to that found within the source domain. The prediction for images at the target domain can be obtained by using M_t directly. Each domain makes use of training, validation and test sets while the labels for the training images in the target domain are withheld.

6.4.2 Learning at Source Domain

Images from the source domain are annotated and the classification of each is independently confirmed by a board-certified pathologist. These images are subsequently used to teach the source domain CNN to map the images into a discriminative feature space. Due to the giga-pixel size of histopathology WSI, each was cropped into manageable sized patches and the cross-entropy loss was adopted $\mathcal{L}_{\mathrm{c}}$ to optimize the performance of the classifier $\mathbf{C}$ in a supervised manner.

$$\mathcal{L}_{\mathrm{c}} = \mathbb{E}_{\mathbf{x}^s \sim \mathcal{S}} - \sum_{i=1}^{N_s} \mathbf{y}_i^s \cdot \log \mathbf{C}(M_s(\mathbf{x}^s; \theta^S)). \tag{6.1}$$

In the above equation, θ^S represents the weights of the source domain CNN. We used a modified fully convolutional AlexNet [112] as the source domain CNN for the classification task. The network does not include a fully connected (FC) layer, instead it only contains convolutional layers. All of the convolutional layers are followed by the Batch Normalization layer [56] and Rectified Linear Units (ReLU), except for the last layer that provides the actual prediction. The details of the network are shown in Figure 6.2A. To achieve the classification for the WSIs, we apply a majority vote on all cropped patches within each WSI which, in turn, provides the prediction.

Due to the high number of domain variations that are exhibited in histopathology

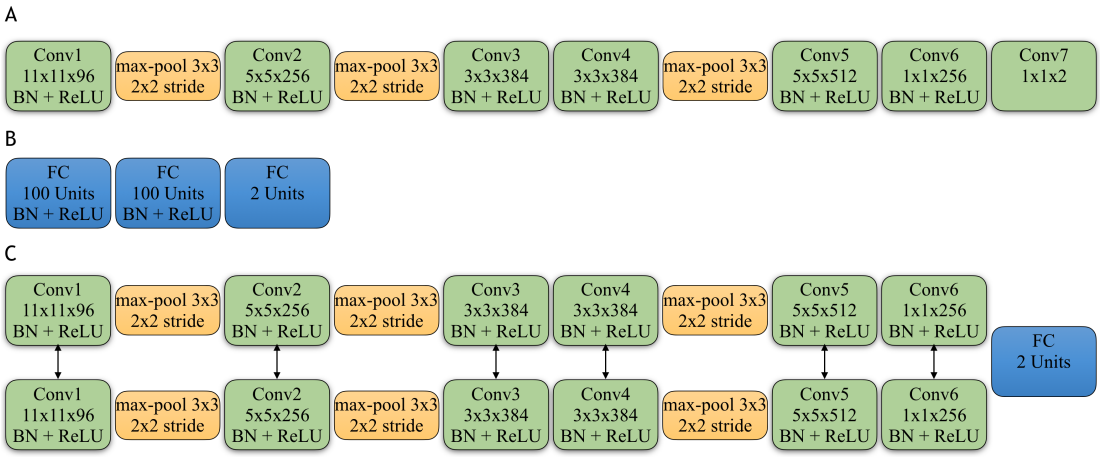


Figure 6.2: Detailed architectures of source domain network, discriminator and Siamese network of target network: (A) The convolutional neural network applied in the source domain. All the convolution layers (Conv) are followed by the Batch Normalization layer (BN) and Rectified Linear Units (ReLU), except for the last Conv layer that gives the classification. The Conv5 and Conv6 layers are also followed by a Dropout layer with the ratio as 0.5. (B)The architecture of the discriminator. All the FC layers are followed by the BN and ReLU, except for the last FC layer that gives the domain prediction. (C) The Siamese network applied in the target domain. The Conv5 and Conv6 layers from the two branches are followed by a Dropout layer with the ratio as 0.5. And the two branches share the same parameters. The feature maps from Conv6 are concatenated to feed into a FC layer to give the similarity prediction between input patches. The Conv6 layers are also followed by a Conv7 layer with the same kernel size as shown in the source domain CNN.

images, the network learned from the source domain may not always generalize sufficiently within the target domain. To address this issue, we introduced two approaches to minimize the domain variations with the details followed.

6.4.3 Color Normalization for Target Domain

The first approach for achieving unsupervised domain adaptation in the histopathology images of target domain utilizes the color normalization. As it can be applied to improve the automated diagnostic performance of histopathology images by decreasing the staining variation among the entire cohort [205, 213, 25, 207].

In order to apply the source mapping M_s on the target domain directly, we transfer the H&E staining information from source domain to the target domain by normalizing

the target images according to a reference image chosen from the source domain. In this case, only the test images from target domain are required to validate the performance while the training images from the target domain are withheld. However, choosing the reference image from source domain is a non-trivial process given the large number of candidate images. Therefore, we uniformly sample a total of N_l reference images from source domain. For each image $\mathbf{x}^t$ in the target domain, we normalize it using each reference image and forward the normalized image $\mathbf{x}_j^t$ into the source domain CNN to generate the logits feature vector. Then we adopt unweighted averaging, as it has been shown as a reasonable ensemble method in deep learning networks [214, 143], to construct the ensemble logits feature $\mathbf{l}_{N_l}$ of $\mathbf{x}^t$ for the N_l iterations, as shown below:

$$\mathbf{l}_{N_l} = \frac{1}{N_l} \sum_{j=1}^{N_l} M_s(\mathbf{x}_j^t; \theta^S). \tag{6.2}$$

Thus the class prediction for $\mathbf{x}^t$ could be achieved by using *softmax* on $\mathbf{l}_{N_l}$. In this study, we apply two color normalization methods, which are Macenko [55] and SPCN [182], as their advantages have been shown in histopathology images [215].

6.4.4 Adversarial Adaptation for Target Domain

The color normalization process makes it possible to perform the stain transfer from source domain to target domain on images directly. The second approach we investigated was unsupervised domain adaptation of histopathology images, in which we explored the adaptation of knowledge on feature space from source to target domain. Therefore we learn a target mapping function M_t, which is a CNN, to map the images from target domain into a discriminate feature space. In order to optimize the target network, we leverage the adversarial training to minimize the discrepancy between the feature space of the target domain and the one of the source domain. We perform asymmetric adaptation where the network at the target domain is fine-tuned from the network of the source domain. Through optimization, the feature space of the target domain learns to mimic the distribution of the source feature space. Thus the target network is trained to extract the domain invariant features from input samples, which

have the same distribution as the source domain. In the process, the training images of target domain are used to carry out the adversarial adaptation.

Adversarial Training

We implement adversarial training following the idea from GAN loss [45] on the feature spaces of source and target domain. The feature vectors generated from the network of source domain or the network of target domain are fed into the discriminator **D**. **D** is trained to map the input feature vectors into a binary domain label, where the "true" denotes the input feature vectors are from source domain and "false" denotes the feature vectors are from target domain. Additionally, the target mapping M_t is learned in an adversarial manner to purposely misdirect the discriminator **D** by reversing the domain label so that the discriminator cannot distinguish between the two feature spaces. Since the mapping parameterization of source model is determined before the adversarial training, we only optimize the target mapping step M_t. By using adversarial learning, we minimize the discrepancy of feature spaces between the source and target domain. Therefore, estimating the category information for the images from target domain can be implemented by M_t. More specifically, the adversarial loss $\mathcal{L}_{\text{adv}_\mathbf{D}}$ for optimizing the discriminator **D** is represented as:

$$\min_{\mathbf{D}} \mathcal{L}_{\text{adv}_\mathbf{D}} = -\mathbb{E}_{\mathbf{x}^s \sim \mathcal{S}} \log \mathbf{D}(M_s(\mathbf{x}^s; \theta^S); \theta^D) - \mathbb{E}_{\mathbf{x}^t \sim \mathcal{T}} \log(1 - \mathbf{D}(M_t(\mathbf{x}^t; \theta^T); \theta^D). \quad (6.3)$$

where θ^T represents the weights of the target domain CNN and θ^D represents the weights of the discriminator. The discriminator is composed of three fully connected layers where each is followed by a Batch Normalization layer and a ReLU layer with the exception of the last one. The details for the architecture of the discriminator are shown in Figure 6.2B. The mapping loss $\mathcal{L}_{\text{adv}_M}$ for optimizing the target mapping M_t is represented as:

$$\min_{M_t} \mathcal{L}_{\text{adv}_M} = -\mathbb{E}_{\mathbf{x}_t \sim \mathcal{T}} \log(\mathbf{D}(M_t(\mathbf{x}^t; \theta^T); \theta^D)). \quad (6.4)$$

For the adversarial training, we optimize the $\mathcal{L}_\text{a}$, where $\mathcal{L}_\text{a} = \mathcal{L}_{\text{adv}_\mathbf{D}} + \mathcal{L}_{\text{adv}_M}$.

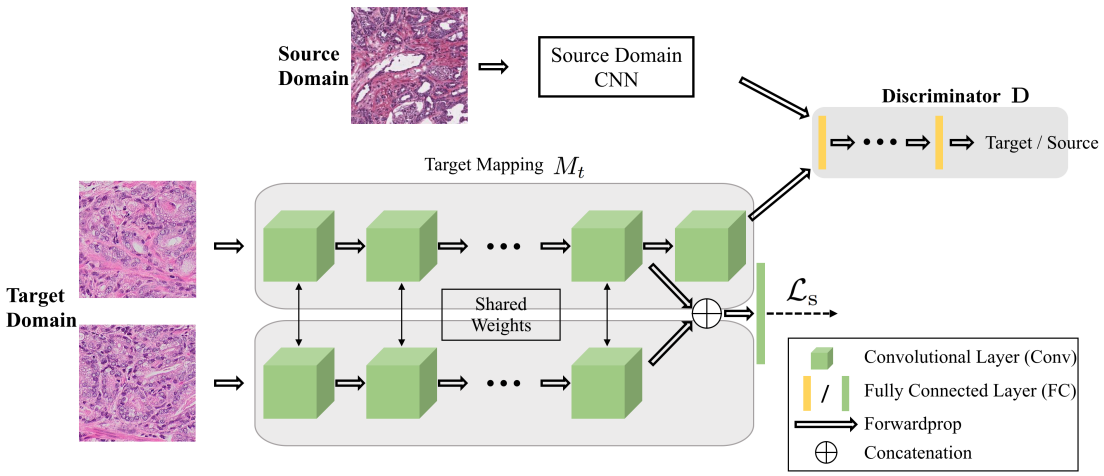


Figure 6.3: The architecture of the networks for the adversarial domain adaptation. The source network and the target network map the input samples into the feature space. The adaptation is accomplished by jointly training the discriminator and target network using the GAN loss to find the domain invariant feature. A Siamese network at target domain adds constrains for patches within the same WSIs.

Siamese Architecture for Target Network

Although there are no annotations for the images at the target domain, the patches cropped from the same WSI should be estimated as the same class by the network at target domain. However, the adversarial loss only forces the distribution of the feature spaces across the two domains to be similar, it can not constrain the target network to determine the similarity of the input samples. Therefore, we introduce a Siamese architecture [216] at target domain to explicitly regularize patches from the same WSI to be classified into the same category. As shown in Figure 6.3, the two identical networks in the target domain share the same weights with the input as a pair of images $(\mathbf{x}_1^t, \mathbf{x}_2^t) \subseteq \mathcal{T} \times \mathcal{T}$. The feature maps obtained from the second to the last layer of the two networks, namely the Conv6 feature maps as shown in Figure 6.2C, are concatenated together to serve as the input vector for a one-layer perceptron to classify the features. Therefore, the input samples are classified by the function $f(\mathbf{x}_1^t, \mathbf{x}_2^t; \theta^F)$, that $f : \mathcal{T} \times \mathcal{T} \mapsto \bar{\mathbf{y}}$ and $\theta^F \subseteq \theta^T$, where $\bar{\mathbf{y}}$=1 indicates input patches belong to the same WSI while $\bar{\mathbf{y}}$=0 denotes not. We learn the binary classifier f using categorical cross-entropy loss $\mathcal{L}_s$ as following:

$$\mathcal{L}_{\mathrm{s}} = \mathbb{E}_{(\mathbf{x}_1^t, \mathbf{x}_2^t) \sim \mathcal{T}} - \sum_{i=1}^{N_p} \bar{\mathbf{y}}_i \cdot \ f(\mathbf{x}_{i1}^t, \mathbf{x}_{i2}^t; \theta^F). \tag{6.5}$$

where N_p denotes the total number of training pairs.

To learn the network at target domain by adversarial adaptation, we adopt a two-stage training process. For the first stage, we train the network at source domain, which is the same as using the color normalization in the adaptation process. For the second stage, we optimize the Siamese network at target domain by applying $\mathcal{L}_{\mathrm{t}}$ where $\mathcal{L}_{\mathrm{t}} = \mathcal{L}_{\mathrm{a}} + \mathcal{L}_{\mathrm{s}}$. For optimizing $\mathcal{L}_{\mathrm{s}}$, we sample the images pairs in the training set of target domain both from the patches cropped from the same WSI and the patches from different WSIs. The learning algorithm for the target network is shown in Algorithm 2.

Algorithm 2: Learning Algorithm for the Network at Target Domain

Input: Initialized target network from source network with weights $\theta^T = \theta^S$

for *number of training iterations* **do**
- sample two same number of mini-batches $\mathbf{x}^s \sim \mathcal{S}$, $\mathbf{x}^t \sim \mathcal{T}$;
- obtain the estimation $\mathbf{y} = M_s(\mathbf{x}^s; \theta^S)$, $\mathbf{y}' = M_t(\mathbf{x}^t; \theta^T)$;
- $\theta^D \leftarrow$ back propagate with stochastic gradient $\nabla \mathcal{L}_{\mathrm{adv_D}}(\mathbf{y}, \mathbf{y}')$;
- $\theta^T \leftarrow$ back propagate with stochastic gradient $\nabla \mathcal{L}_{\mathrm{adv_M}}(\mathbf{y}')$;
- sample mini-batches with paired of images $\mathbf{x}_1^t, \mathbf{x}_2^t \sim \mathcal{T}$;
- obtain the estimation $\bar{\mathbf{y}} = f(\mathbf{x}_1^t, \mathbf{x}_2^t; \theta^F)$;
- $\theta^F \leftarrow$ back propagate with stochastic gradient $\nabla \mathcal{L}_{\mathrm{s}}(\bar{\mathbf{y}})$;

6.5 Experiments

In this section, we validate the proposed approaches using the unsupervised domain adaptation for the classification of the histopathology images.

6.5.1 Implementation Details

We conducted two sets of unsupervised domain adaptation for classification of prostate histpathology images, which are TCGA (w/o UP) $\rightarrow$ UP and TCGA $\rightarrow$ RCINJ. We firstly use the images in source domain to train a binary classification network. The data from source domain is randomly divided into the training and the testing sets at a ratio of 80% (validation set is randomly selected from the training set) / 20%. The

patients with more than one WSI can only contribute the images to the training set or the testing set. During the training process, the images are resized as 256×256 and randomly cropped to 224×224 to feed into the network. During the testing process, all the patches are resized to 256×256, we do the single center-crop for all testing patches. The network is trained from scratch. For the adaptation using color normalization, we utilize the source domain CNN as the network for target domain to determine the prediction from the testing set. For the adversarial adaptation, we optimize the Siamese network at target domain by fixing the parameters of source domain CNN and training the target network and the discriminator network at the same time. The prostate images at the target domain are randomly divided into the training and the testing sets at a ratio of 80% and 20%.

Our implementation is based on Tensorflow [217]. To train the source network, we use mini-batch Stochastic Gradient Descent (SGD) with mini-batch size as 128. The momentum is 0.9 and the weight decay is 0.0005. The initial learning rate is 0.001 and periodically annealed by 0.1. To train the target network for the adversarial adaptation, we use Adam optimization [218] with the fixed learning rate as 0.00001. The mini-batch size for optimizing $\mathcal{L}_{\mathrm{a}}$ and $\mathcal{L}_{\mathrm{s}}$ is set as 128.

6.5.2 Source Domain Performance

As the training process contains two steps, we first show the performance of the network at the source domain. The comparison between the source network and the previous study [188] is shown in Table 6.3. From the results, we can see both of our models have better performance than [188]. However, the study at [188] uses less WSIs than ours and the network with the best performance reported in [188] is wider and deeper than our study. Although such differences lead to biased comparison, it could still demonstrate the source domain network is well trained to classify the TCGA prostate images into low Gleason score and high Gleason score. We have tried deeper network, such as ResNet-50 [143], but the modified AlexNet used in the study has a better performance. For example, the modified AlexNet has the accuracy of 83.0% on TCGA while the ResNet-50 [143] has the accuracy as 79.8%.

Table 6.3: The source domain network performance. The source domain classification network outperforms previous study [188] using prostate cancer data from TCGA without UP and TCGA. The source domain network using one all TCGA prostate cancer data achieves higher classification accuracy than using TCGA without UP because of more data included for training the network.

	Accuracy (%)
Previous Study [188]	73.5
TCGA (w/o UP)	**76.9**
TCGA	**83.0**

6.5.3 Comparison Results

In this section, we show the comparative results using different approaches for learning the classification model at the target domain.

Adaptation using Color Normalization

First, we show the domain adaptation results only using color normalization. The qualitative results for the color normalization are shown in Figure 6.4. We sample different number of reference images, which is N_l in Equation 6.2, due to the large number of training set in source domain. For each color normalization method, we use N_l-Ensemble to indicate the number of reference images. For each N_l, we run the experiments for 10 times and report the mean and the standard deviation values in Table 6.4. Additionally, we show the baseline results in Table 6.4 where the source domain CNN is applied on the original images from target directly. We can see that due to the different image distributions of the source and target domains, the network learned from source domain is not working appropriately when applied on target domain directly. For the adaptation of TCGA (w/o UP) $\rightarrow$ UP, the results show using the two color normalization methods both improve the classification accuracy and with more reference images, it could achieve the better classification. Furthermore, SPCN [182] achieves better results compared to Macenko [55] as it has higher mean classification accuracy and less standard deviation. While for the adaptation of TCGA $\rightarrow$ RCINJ, no better result is observed by using the color normalization, which indicates color normalization may not be robust when applied for the domain adaptation of the prostate

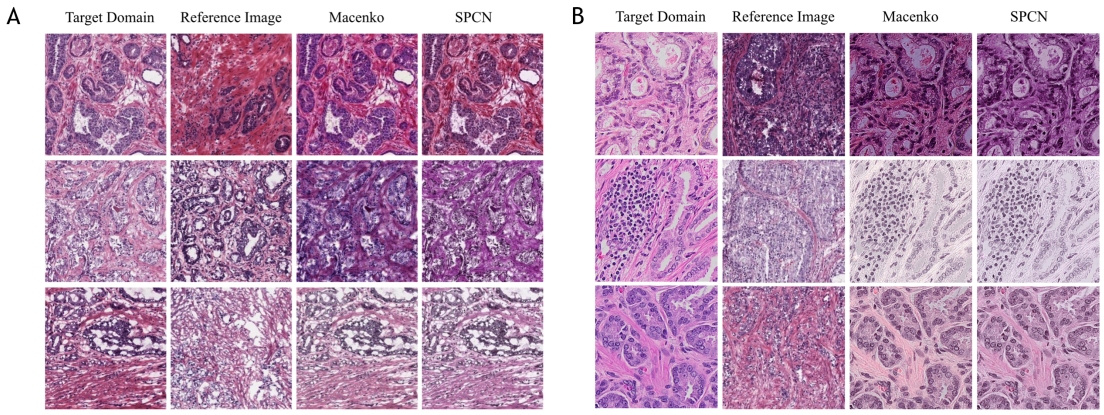


Figure 6.4: Example images selected from the testing set of target domain are normalized by the reference images sampled from the training set of source domain using two color normalization methods including Macenko [55] and SPCN [182]. (A) The adaptation of TCGA (w/o UP) → UP. (B) The adaption of TCGA → RCINJ.

histopathology images. For both TCGA (w/o UP) → UP and TCGA → RCINJ, using more reference images could decrease the standard deviation of the ensemble results. On the other hand, the high standard deviation indicates the high sensitivity when choosing a reference image, which makes the color normalization less practicable for unsupervised domain adaptation given the difficulty of deciding the optimal reference image within the source domain.

Additionally, we show the comparison with color augmentation, which has been proved effective for the data augmentation of histopathology images [220, 221, 219]. We follow the methods introduced in [219] where random color perturbations is applied on each patch in the training set. Experimental results in Table 6.4 show the color augmentation is more effective than color normalization on the two sets of experiments.

Adversarial Adaptation

Second, we show the results of using the adversarial domain adaptation for TCGA (w/o UP) → UP and TCGA → RCINJ. The quantitative results for the adaptation are shown in Table 6.4. Through the adversarial adaptation, we could effectively adopt the discriminative knowledge from TCGA (w/o UP) to the UP and from TCGA to

Table 6.4: Unsupervised domain adaptation for TCGA (w/o UP) → UP and TCGA → RCINJ using color normalization and adversarial adaptation. The classification accuracy of two color normalization methods including Macenko [55] and SPCN [182] with different number of ensembles, and the target network with adversarial loss ($\mathcal{L}_{\text{a}}$) only and the target network with adversarial loass and Siamese loss together ($\mathcal{L}_{\text{t}}$) are shown for two sets of adaptations. We also compare our approach with color augmentation [219]. Our proposed approach has a better performance than other state-of-the-art study [189] on the unsupervised adaptation task.

	TCGA (w/o UP) → UP	TCGA → RCINJ
Baseline	54.3	56.3
Macenko [55] 1-Ensemble	65.7 ± 11.9	51.3 ±6.1
Macenko [55] 2-Ensemble	70.0 ± 5.9	53.8 ±8.5
Macenko [55] 5-Ensemble	72.3 ± 3.8	55.0 ± 7.3
Macenko [55] 10-Ensemble	72.6 ± 2.3	55.0 ± 4.7
SPCN [182] 1-Ensemble	70.0 ± 7.3	56.3 ± 13.4
SPCN [182] 2-Ensemble	71.7 ± 6.7	55.0 ± 15.3
SPCN [182] 5-Ensemble	72.9 ± 2.6	55.6 ± 9.8
SPCN [182] 10-Ensemble	73.4 ± 1.8	54.4 ± 8.4
Color Augmentation [219]	74.5	56.3
Generate-to-Adapt [189]	71.7	62.5
$\mathcal{L}_{\text{a}}$ only	71.4± 1.1	62.5 ± 2.5
$\mathcal{L}_{\text{t}}$	**77.1± 1.1**	**75.0 ± 2.5**

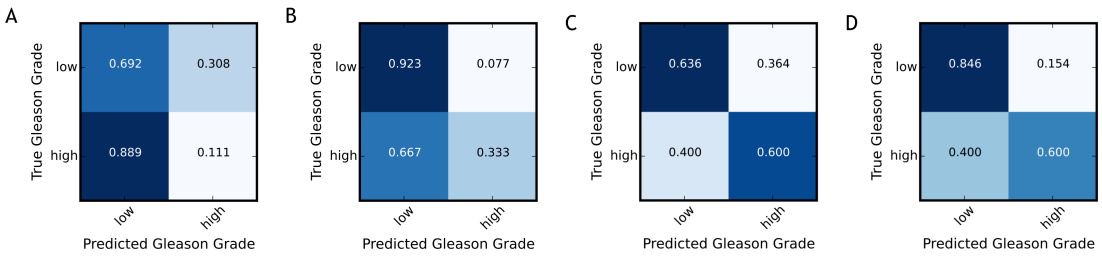


Figure 6.5: The confusion matrix of the target network before and after the adaptation for TCGA (w/o UP) → UP and TCGA → RCINJ. (A) The confusion matrix for UP before domain adaptation. (B) The confusion matrix for UP after domain adaptation. (C) The confusion matrix for RCINJ before domain adaptation. (D) The confusion matrix for RCINJ after domain adaptation.

RCINJ without requiring additional annotations. Compared with the adaptation using color normalization, the adversarial adaptation achieves better classification results for the two setting of experiments, which demonstrates its effectiveness and robustness. Additionally, we compare our approach with the Generate-to-Adapt [189] on the two tasks and our approach outperforms the current, state-of-the-art algorithm of the unsupervised domain adaptation.

We further calculate the statistically significance of the accuracy improvement between the adapted network and the baseline network using McNemar Test [222] and demonstrates the improvement of classification accuracy is statistically significant with a p-value less than 0.05. In addition, we show the result of the ablation study in Table 6.4 that using $\mathcal{L}_{\mathrm{t}}$ achieves better classification accuracy than $\mathcal{L}_{\mathrm{a}}$ only. Figure 6.5A-6.5B show the confusion matrices for the adaptation for TCGA (w/o UP) → UP and Figure 6.5A-6.5B show the confusion matrices for the adaptation of TCGA → RCINJ. Compared to before domain adaptation and after domain adaptation, the true low-grade classification accuracy are significantly improved. It is crucial for prostate cancer diagnosis for patients with low Gleason grade is one of the main criteria for active surveillance and intervention.

We show the qualitative results for TCGA → RCINJ in Figure 6.6. We use the probability predicted by the network on the patches to generate a classification probability heatmap and overlay the heatmap on the original image. The red color indicates the

high Gleason score and blue color indicates the low Gleason score. Figure 6.6A-B show example prostate WSIs from RCINJ with the low Gleason score and the ground-truth heatmap overlaid on it. Figure 6.6C shows the WSI with high Gleason score. After the unsupervised domain adaptation, the target network could correctly classify most of patches into the correct Gleason score.

6.6 Discussion and Conclusion

In this paper, we investigate viable approaches for addressing the challenges presented by the heterogeneous characteristics exhibited within digitized specimens, that arises when analyzing samples that have been prepared at disparate laboratories and institutes. We present two different unsupervised domain adaptation methods to resolve the domain variations to make it possible to render accurate classification of imaged histopathology specimens. To meet the requirements of this endeavor required color normalization to transfer the staining information from images in source domain to the images in target domain whereas adversarial training was implemented to transfer the discriminate information in feature space from the source to the target domain. Throughout these experiments, our team utilized a well trained CNN at source domain that was shown to outperform other methods used on the TCGA prostate cancer dataset. This work shows that when compared with color normalization, adversarial training is more robust for performing unsupervised domain adaptation, indicating that adversarial training may also serve to decrease the differences in the morphologic and structural patterns for histopathology images that can be introduced during processing at disparate institutions. In this research, we further proposed to leverage a Siamese architecture to add the regularization for the target domain to achieve better results than that resulting from utilizing the state-of-the-art method for unsupervised domain adaptation. Due to the limited size of the datasets in these feasibility studies, we plan to conduct expanded experiments using a wider range of histopathology image classification problems.

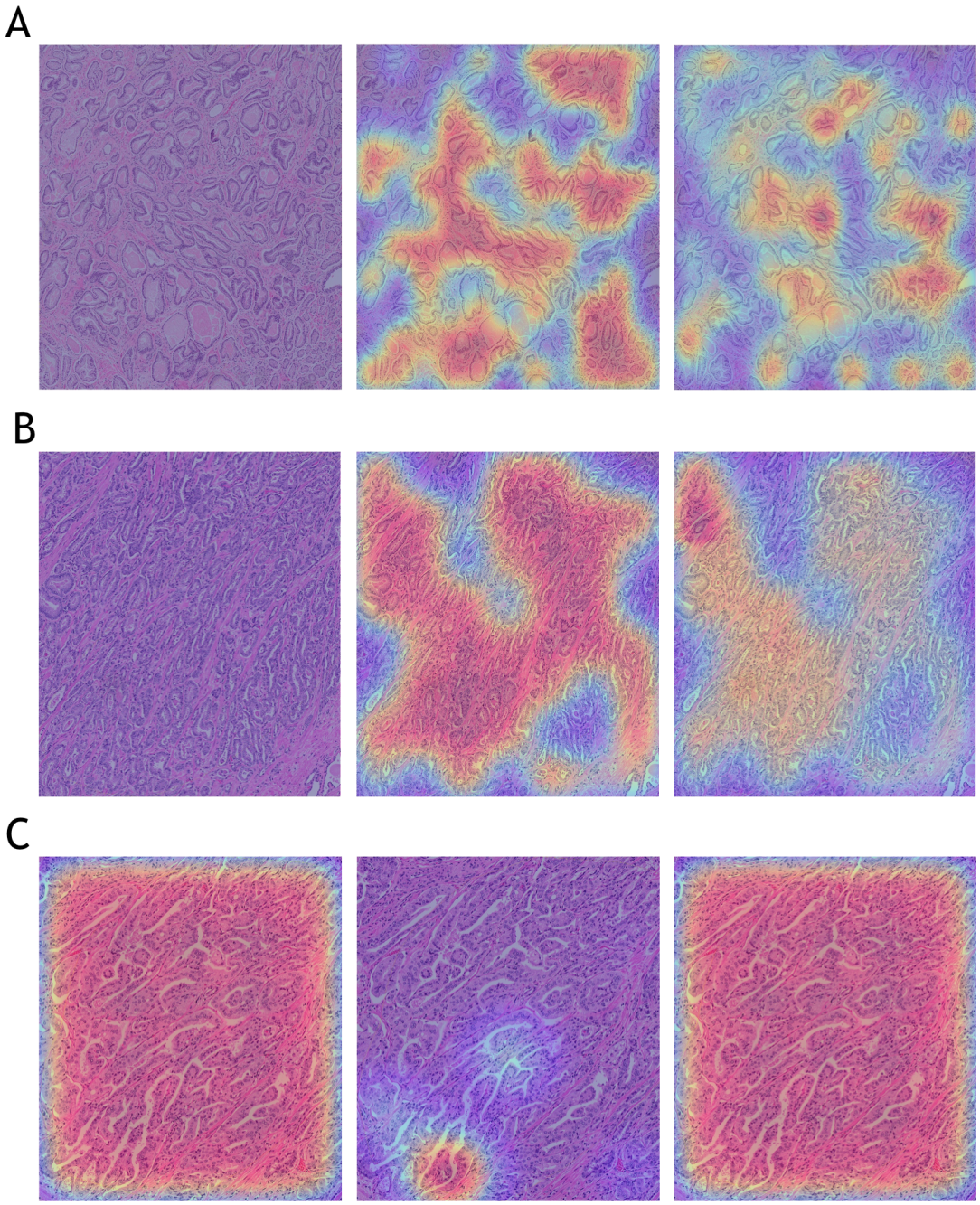


Figure 6.6: (A) and (B) show the example images from RCINJ with Gleason score 6. (C) shows the example image from RCINJ with Gleason score 8. The left column shows the original images with heatmaps overlaid on them; the middle column shows the heatmaps generated from the baseline model (using source domain network); the right column shows the heatmaps generated from the model optimized by $\mathcal{L}_{\mathrm{t}}$.

Chapter 7

Nuclei Detection Ensemble Workflows Across Clustered Infrastructure

7.1 Introduction

In this part, we propose a new approach to parallelize the nuclei detection algorithm by utilizing CometCloud to speed up the whole process to make the nuclei segmentation running in real-time a possibility.

Diseases such as cancer can cause changes in tissue morphology at the sub-cellular levels. The shape and texture properties of nuclei and changes in these properties provide diagnostic value to determine disease stage and are sources of rich information with which to study disease biology. Traditionally pathologists have examined tissues under high power microscopes. Advances in digital microscopy technologies have enabled high-resolution images from whole slide tissues. Analyses of tissue images allow for a quantitative assessment of nuclear morphology and can lead to a better understanding of the mechanisms of disease onset and progress and to better strategies for curing disease. The Cancer Genome Atlas (TCGA) project, for example, has collected about 30,000 whole slide tissue images from over 25 cancer types. In this work we use tissue images from the TCGA repository. While tissue images contain rich morphological information, the extraction of this information (via segmentation of nuclei and computation of shape and texture features) is a computationally challenging problem.

A robust nuclei segmentation algorithm has been reported in [35], which includes two main sequential steps, seed detection and contour generation. Within the medical images, some nuclei are isolated ones; some are overlapping with each other. It's not very efficient to run our previous nuclei segmentation algorithm directly on the whole image which may contains hundreds and thousands of nuclei. To accelerate the

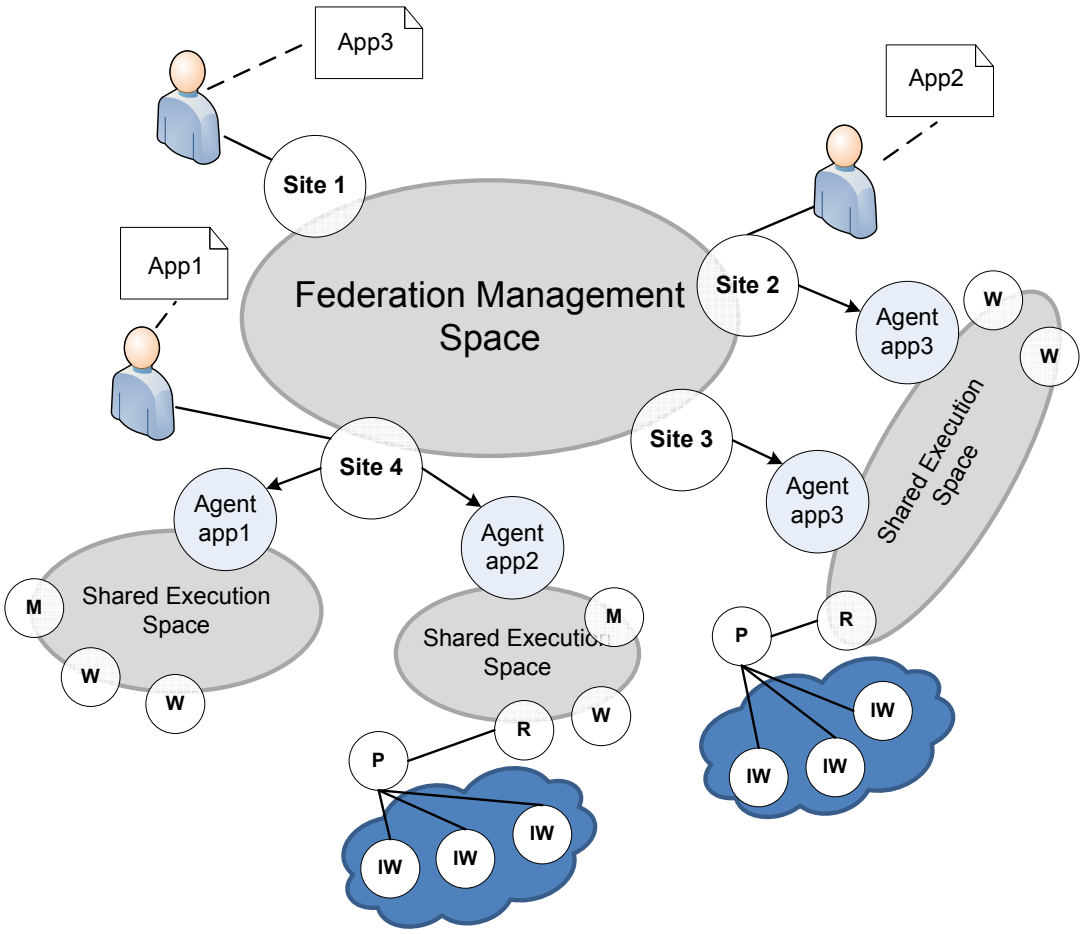


Figure 7.1: CometCloud Federation Model.

process, there have been many applications using cloud computing on medical image analysis, but most of them were focused on data parallelization instead of the algorithm parallelization [36] [37] [38] [39]. Our earlier work focused on fluorescence images while this work addresses the challenge of working with specimens which have not been enhanced with specialized staining methods and can be used across a broader number of application areas.

7.2 CometCloud

CometCloud is an autonomic framework designed to enable highly heterogeneous, dynamically federated computing and data platforms that can support end-to-end application workflows with diverse and dynamic changing requirements. CometCloud provides

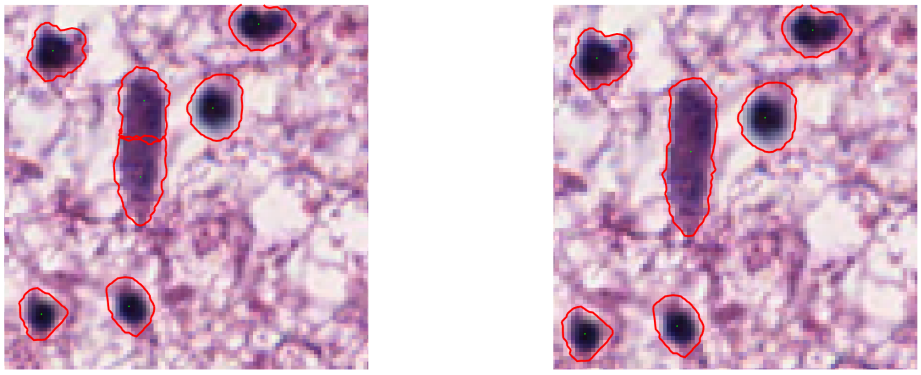
interfaces to independently describe application workflows and resources. Application workflows are currently described using XML documents, as a set of stages defining input and output data, dependencies to other stages, scheduling policies, and possibly annotated with specific objectives and policies.

Objectives and polices leverage CometCloud autonomic mechanisms to drive resource provisioning and execution of application workflows while satisfying user constraints (e.g., budget, deadline) and application requirements (e.g., type of resources). The autonomic mechanisms in place not only provision the right resources when needed, but also monitor the progress of the execution and adapt the execution to prevent violations of established agreements [223].

CometCloud uses a federation approach to aggregate heterogeneous and geographically distributed resources. These resources are exposed to users as a seamless elastic pool of resources. The CometCloud federation, illustrated in Figure 7.1, is created dynamically and collaboratively, where resources/sites can join or leave at any point, identify themselves (using security mechanisms such as public/private keys), negotiate terms of federation, discover available resources, and advertise their own resources and capabilities [224]. This federation is coordinated using tuple-spaces, called CometSpaces [225]. Specifically, we define two types of coordination spaces. First, a single management space spans across all resource sites creating and orchestrating the federation. Second, multiple shared execution spaces are created on-demand during application workflow executions to satisfy computational or data needs. Execution spaces can be created within a single resource site, or can burst to others, such as public clouds or external HPC systems.

7.3 Enabling nuclei detection workflow on CometCloud

Running the large image dataset on CometCloud means that we can't adjust the parameters of nuclei detection for each image, so it's important to set the parameters for each image automatically. Also, in order to increase the efficiency and parallelize the algorithm, each image needs to be divided into many sub-images, here are the nuclei



(a) (b)

Figure 7.2: (a) is the image with false multiple seeds within one single nucleus (b) is the image with merging seeds on the elongated shape nuclei. The red lines indicate nuclei contour and the green dots indicate seed in the nuclei.

region images. And the sub-images can be sent to each agent to implement nuclei detection algorithm.

7.3.1 Parallelizing each image into multiple nuclei region images

Each image contains many nuclei. To parallelize the nuclei segmentation process, each image is divided into multiple sub-images, within which each contains individual nuclear region. Because the nuclei in LGG and GBM are more similar than the nuclei in COAD, LUAD and PAAD, we use two different pre-processing algorithms on them. For the LGG and GBM images, the color of nuclei is almost homogenous and solid; while for the other three data sets, there are white areas in the nuclei and the boundary of nuclei is not clear.

Pre-processing for LGG and GBM

For the LGG and GBM images, we use color deconvolution to get stain vector that includes the nuclei regions [55]. Then a small structure element is used to reconstruct the image to remove the small connected objects. The shape of the structure is a disk shape since it's very similiar to the shape of the nuclei. The reconstruction process is morphological opening the image followed by morphological closing the image. Because

the image after color deconvolution may contain small objects that connect different nuclei regions, the reconstruction process can remove these small objects and keep each nuclei region in whole. Using multi-class SVM classifier, a threshold is predicted. Training image feature from LGG and GBM datasets is the index image with 255 color vectors. Using the predicted threshold, a binary image is extracted accordingly.

Pre-processing for COAD, LUAD and PAAD

For the COAD, LUAD and PAAD images, we also use color deconvolution to get stain vector that includes the nuclei region. But the image reconstruction process is different from the process of LGG and GBM. Because there are white areas in the nuclei region and it leads to one nuclei area may be separated into two parts after color deconvolution. So we reconstruct the image by morphological closing and opening with a small structure element. The shape of the structure element is also a disk shape. Therefore if one nuclei is separated into two parts, they can be connected again and also the small objects will be removed. Then we use Fast Radial Symmetry Transform(FRST) on the reconstructed image to get the binary image [226].

Parallelizing the nuclei segmentation

After having the binary image from pre-processing, we use the nuclei regions to calculate the average diameter in each image, which is used for seed detection step within nuclei segmentation. The nuclei regions whose areas are less than 80^{th} percentile of the whole image and the whose ratio of the maximum length to minimum length less than 1.3 are chosen as the regions with only one nuclei inside. The number of nuclei regions equals to the number of the newly generated images, denoted as sub-images of the original one. The nuclei region images can contain only one single nucleus or multiple overlapping nuclei. The seed detection is implemented directly on those nuclei region images followed by their corresponding contour generation.

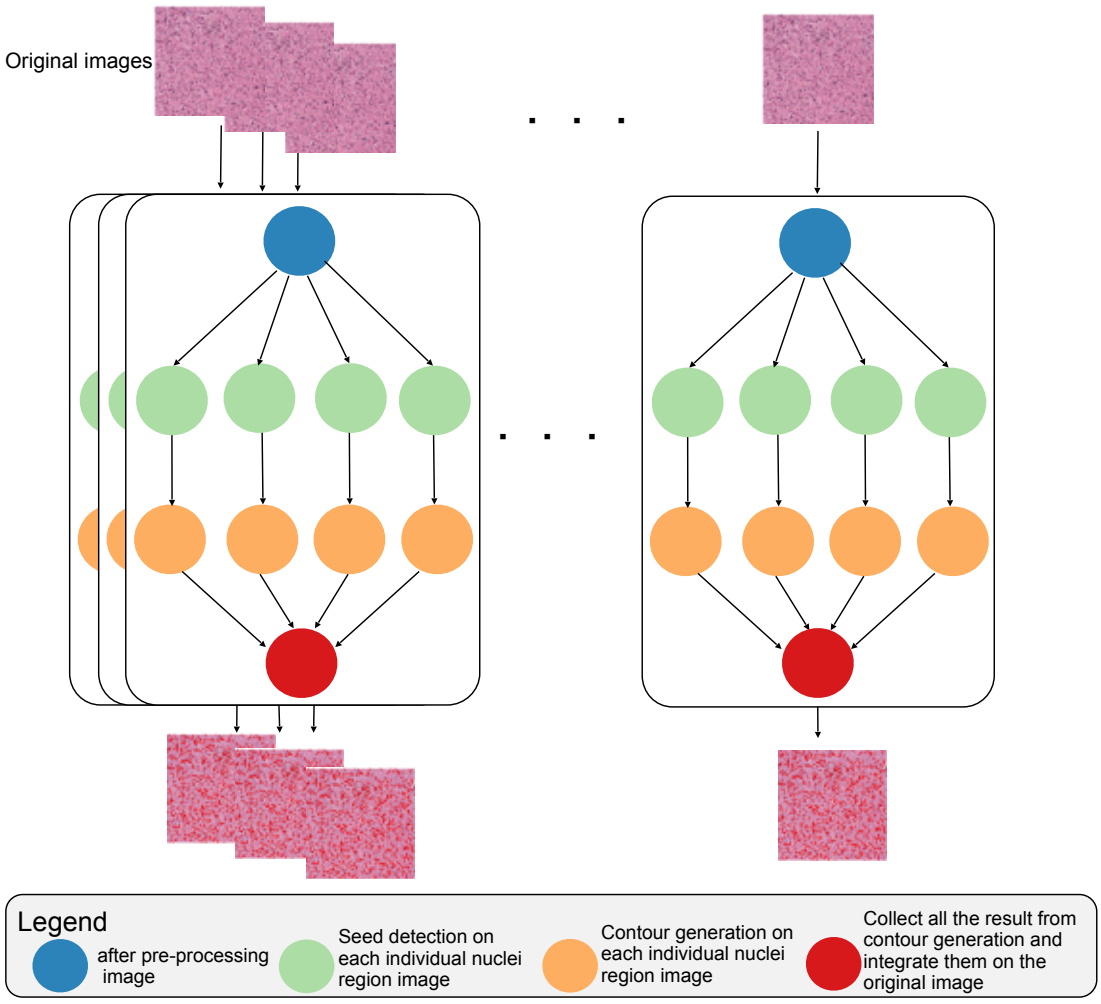


Figure 7.3: Nuclei Detection Workflow

Merging seeds on elongated nuclei

For the elongated nuclei, our previous seed detection method may create multiple false seeds inside the single nucleus. To address this issue, merging the seeds on elongated nuclei is proposed. If the nuclei region has concave points and the standard deviation of gray image of the nuclei region is less than the standard deviation of gray image of all the nuclei region, the multiple seeds within the nuclei region are merged by averaging the coordinates of all the seeds to create a single one, as show in Figure 7.2.

Table 7.1: Average running time comparison.

	LGG	GBM	COAD	LUAD	PAAD
Average running time with CometCloud for each image (seconds)	2013	5348	5228	3444	6210
Average running time with only pre-processing for each image (seconds)	3318	8682	10636	6656	8068
Average running time with previous algorithm for each image (seconds)	6754	7885	11728	9920	8340

7.3.2 Workflow on CometCloud

There are three sequential steps for the workflow running on CometCloud, as shown in Figure 7.3. Before running step 1, all the original images are read from a directory and pre-processing is applied for the two main types of images as indicated above. Within Step 1, the processed images are divided into multiple nuclei region images. Within step 1, because there are different number of nuclei including isolated and overlapping ones within each image, different number of nuclei region images are created for a single original image. Within step 2, nuclei detection algorithm is implemented on each nuclei region image which is treated as a single task. It includes two sequential tasks, seed detection and contour generation on each individual nuclei region image. After finishing step 2, step 3 is to collect all the result from contour generation and integrate them on the original image. Figure 7.4 shows five example images after running nuclei detection with CometCloud.

7.4 Experiment Results

We use five different datasets from TCGA, which include LGG, GBM, COAD, LUAD and PAAD. They include 25 LGG images, 15 GBM images, 15 COAD images, 10 LUAD images and 10 PAAD images. Those images are all 1024×1024 and under 20× objective. The new parallelization algorithm and the previous algorithm are both tested on the same machine. The whole process is implemented on clusters with Intel Xeon

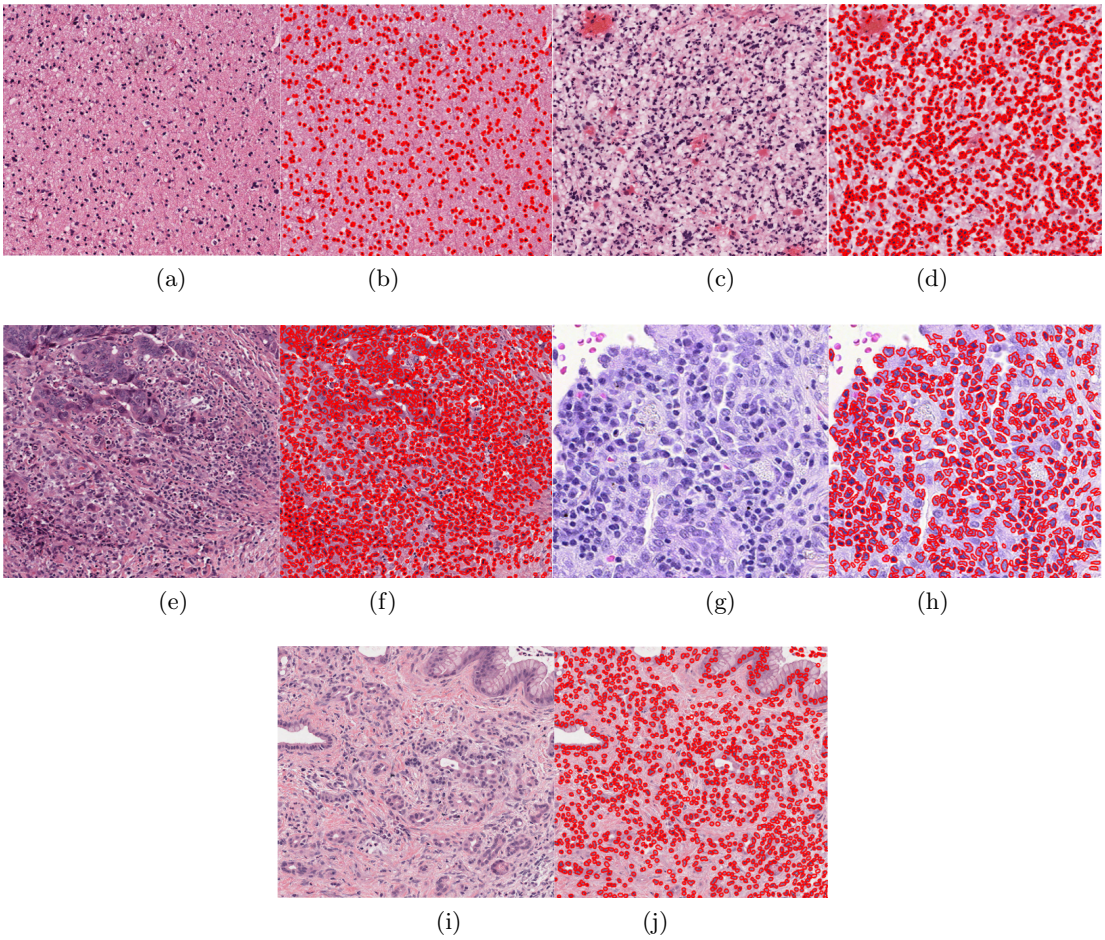


Figure 7.4: (a) and (b) are LGG images, (c) and (d) are GBM images, (e) and (f) are COAD images, (g) and (h) are LUAD images, (i) and (j) are PAAD images. (a), (c), (e), (g) and (i) are original images and (b), (d), (f), (h) and (j) are the images with nuclei detected using CometCloud.

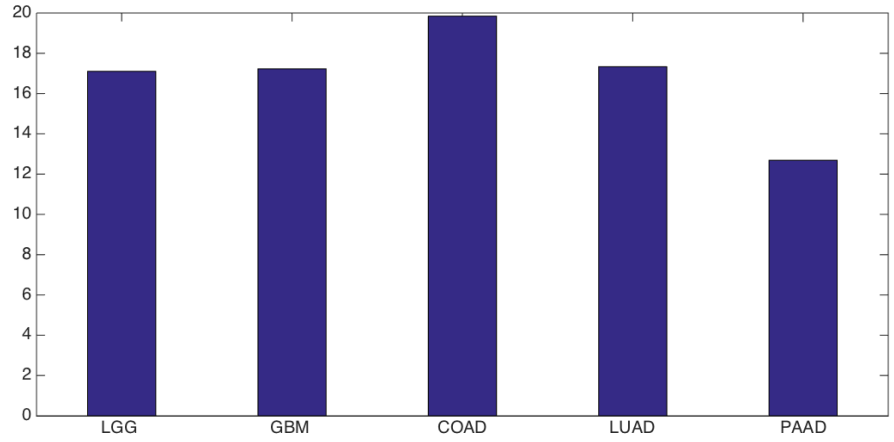


Figure 7.5: Average speedup ratio of with and without using parallelization on CometCloud using 32 machines

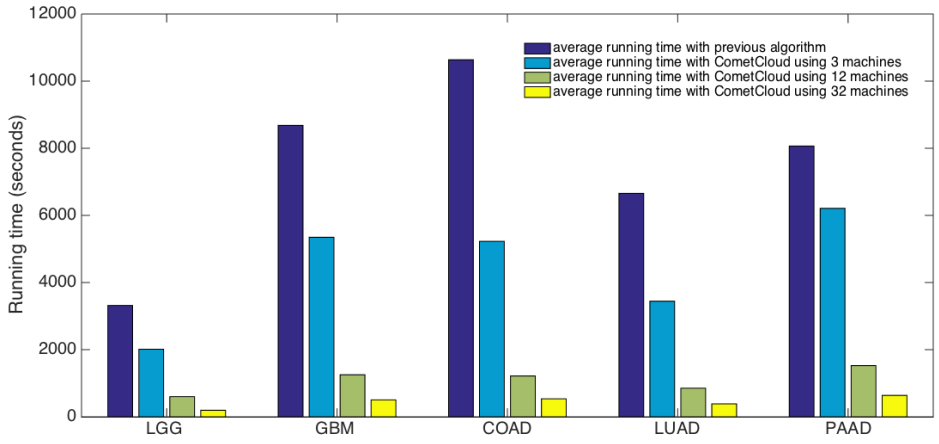


Figure 7.6: Average running time comparison of previous algorithm and using CometCloud with different number of machines

E56620 of 8 cores and 24GB memory each. The algorithms are written with Octave 3.4.3 and used Octave image package 1.0.15. For step 1 and step 3 in the workflow, there are 25, 15, 15, 10 and 10 tasks for LGG, GBM, COAD, LUAD and PAAD images respectively. Because the number of tasks equals to the number of images in the two steps. And for step 2 in the workflow, there are 6830, 9823, 6960, 4727 and 8827 tasks for LGG, GBM, COAD, LUAD and PAAD images respectively. Because step 1 generated thousands of images for each dataset, so each image is an individual task and the number of tasks are different for different types of dataset. The results are shown in Figure 7.5 and Table 7.1. Table 7.1 denotes the average running time for various image types. Figure 7.5 represents the average speedup ratio of with and without using parallelization on CometCloud using 32 machines. Also, the running time decreases significantly if we increase the number of machines for agents, as shown in Figure 7.6.

7.5 Conclusion

In this work, we develop a new parallelization nuclei segmentation algorithm based on CometCloud. The algorithm is tested on five types of TCGA datasets. The running time could be significantly decreased by using this new parallelization algorithm compared with the previous nuclei segmentation algorithm. From the results, by increasing the number of machines for agent, nuclei detection in the workflow takes much less time, that makes the nuclei segmentation running in real-time in practice. Meanwhile this work addresses the challenge of working with specimens which have not been enhanced with specialized staining methods and it can be used across a broader number of application areas.

Chapter 8

Discussion

In conclusion, the thesis introduces robust computer-aided methods to have a better diagnosis and prognosis for prostate cancer, especially for Gleason pattern 3 and 4. We propose a method for quantitatively analyzing histopathology prostate cancer images by segmenting glandular regions and grade those regions. The proposed approach may lead to a more reliable method to assist pathologists in performing the stratification of prostate cancer patients and improves therapy planning. Furthermore, we study how to quantify image features from histopathology images and use the features for recurrence analyses on different survival models. Additionally, we introduce an effective way to combine histopathology images and genomic features to obtain computational biomarkers, which are more closely correlated with patients' recurrence risk compared to standard clinical prognostic factors and engineered image texture features. The results of our study suggest that these approaches could be utilized to predict recurrence and progression for patients with prostate cancer. Moreover, to have robust models, we study viable approaches for addressing the challenges presented by the heterogeneous characteristics exhibited within digitized specimens, that arise when analyzing samples that prepared at different laboratories and institutes. We show the introduced method is robust for performing unsupervised domain adaptation, indicating it may also serve to decrease the differences in the morphologic and structural patterns for histopathology images that can be introduced during processing at different institutions. Finally, we study how to develop more efficient algorithms and propose a new parallelization nuclei segmentation algorithm based on CometCloud. The algorithm is tested on five types of TCGA datasets. The running time can be significantly decreased by using this new parallelization algorithm compared with the previous nuclei segmentation algorithm.

In the following, we show several potential future directions that can be based on our current studies and can be serverd on a wider purpose.

Searching Neural Networks Although CNN has gained tremendous success in recent years, choosing or designing neural networks for analyzing histopathology images is still a challenging task, especially there are many different kinds of cancer images. Recent studies have shown the advantage of automatically searching CNN [227]. Searching CNN can be applied on the analysis of histopathology images that new networks can be found to achieve better performance on different datasets.

Generating New Data Currently, preparing and labeling histopathology images are still time and money consuming. Also, getting the genomic information from patients is very expensive that not everyone could afford. The algorithms that could synthesize genomic data given histopathology images will benefit the community a lot because it can save money on obtaining genomic data from patients and the combination of genomic data and histopathology images can provide better diagnosis information.

References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2017," *CA: a cancer journal for clinicians*, vol. 67, no. 1, pp. 7–30, 2017.

[2] F. H. Schröder, J. Hugosson, M. J. Roobol, T. L. Tammela, S. Ciatto, V. Nelen, M. Kwiatkowski, M. Lujan, H. Lilja, M. Zappa *et al.*, "Prostate-cancer mortality at 11 years of follow-up," *New England Journal of Medicine*, vol. 366, no. 11, pp. 981–990, 2012.

[3] A. Wolf, R. C. Wender, R. B. Etzioni, I. M. Thompson, A. V. D'Amico, R. J. Volk, D. D. Brooks, C. Dash, I. Guessous, K. Andrews *et al.*, "American cancer society guideline for the early detection of prostate cancer: update 2010," *CA: a cancer journal for clinicians*, vol. 60, no. 2, pp. 70–98, 2010.

[4] P. M. Pierorazio, P. C. Walsh, A. W. Partin, and J. I. Epstein, "Prognostic gleason grade grouping: data based on the modified gleason scoring system," *BJU international*, vol. 111, no. 5, pp. 753–760, 2013.

[5] L. Egevad, T. Granfors, L. Karlberg, A. Bergh, and P. Stattin, "Prognostic value of the gleason score in prostate cancer," *BJU international*, vol. 89, no. 6, pp. 538–542, 2002.

[6] D. F. Gleason, G. T. Mellinger, L. J. Arduino, J. C. Bailar III, L. E. Becker III, H. I. Berman III, A. J. Bischoff, D. P. Byar, C. E. Blackard, R. P. Doe *et al.*, "Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging," *The Journal of urology*, vol. 111, no. 1, pp. 58–64, 1974.

[7] J. I. Epstein, A. W. Partin, J. Sauvageot, and P. C. Walsh, "Prediction of progression following radical prostatectomy: a multivariate analysis of 721 men with long-term follow-up," *The American journal of surgical pathology*, vol. 20, no. 3, pp. 286–292, 1996.

[8] A. Billis, M. S. Guimaraes, L. L. Freitas, L. Meirelles, L. A. Magna, and U. Ferreira, "The impact of the 2005 international society of urological pathology consensus conference on standard gleason grading of prostatic carcinoma in needle biopsies," *The Journal of urology*, vol. 180, no. 2, pp. 548–553, 2008.

[9] C.-C. Pan, S. R. Potter, A. W. Partin, and J. I. Epstein, "The prognostic significance of tertiary gleason patterns of higher grade in radical prostatectomy specimens: a proposal to modify the gleason grading system," *The American journal of surgical pathology*, vol. 24, no. 4, pp. 563–569, 2000.

[10] A. Freeman, "Prognostic gleason grade grouping: data based on the modified gleason scoring system," *BJU international*, vol. 111, no. 5, pp. 691–692, 2013.

[11] T. Y. Chan, A. W. Partin, P. C. Walsh, and J. I. Epstein, "Prognostic significance of gleason score 3+ 4 versus gleason score 4+ 3 tumor at radical prostatectomy," *Urology*, vol. 56, no. 5, pp. 823–827, 2000.

[12] W. A. Sakr, M. V. Tefilli, D. J. Grignon, M. Banerjee, J. Dey, E. L. Gheiler, R. Tiguert, I. J. Powell, and D. P. Wood, "Gleason score 7 prostate cancer: a heterogeneous entity? correlation with pathologic parameters and disease-free survival," *Urology*, vol. 56, no. 5, pp. 730–734, 2000.

[13] W. K. Lau, M. L. Blute, D. G. Bostwick, A. L. Weaver, T. J. Sebo, and H. Zincke, "Prognostic factors for survival of patients with pathological gleason score 7 prostate cancer: differences in outcome between primary gleason grades 3 and 4," *The Journal of urology*, vol. 166, no. 5, pp. 1692–1697, 2001.

[14] C. Herman, M. Kattan, M. Ohori, P. Scardino, and T. Wheeler, "Primary gleason pattern as a predictor of disease progression in gleason score 7 prostate cancer: a multivariate analysis of 823 men treated with radical prostatectomy," *The American journal of surgical pathology*, vol. 25, no. 5, pp. 657–660, 2001.

[15] K. K. Rasiah, P. D. Stricker, A.-M. Haynes, W. Delprado, J. J. Turner, D. Golovsky, P. C. Brenner, R. Kooner, G. F. O'Neill, J. J. Grygiel *et al.*, "Prognostic significance of gleason pattern in patients with gleason score 7 prostate carcinoma," *Cancer*, vol. 98, no. 12, pp. 2560–2565, 2003.

[16] D. V. Makarov, H. Sanderson, A. W. Partin, and J. I. Epstein, "Gleason score 7 prostate cancer on needle biopsy: Is the prognostic difference in gleason scores 4 3 and 3 4 independent of the number of involved cores?" *The Journal of urology*, vol. 167, no. 6, pp. 2440–2442, 2002.

[17] J. R. Stark, S. Perner, M. J. Stampfer, J. A. Sinnott, S. Finn, A. S. Eisenstein, J. Ma, M. Fiorentino, T. Kurth, M. Loda *et al.*, "Gleason score and lethal prostate cancer: does 3+ 4= 4+ 3?" *Journal of Clinical Oncology*, vol. 27, no. 21, pp. 3459–3464, 2009.

[18] S. M. Khoddami, S. F. Shariat, Y. Lotan, H. Saboorian, J. D. McConnell, A. I. Sagalowsky, C. G. Roehrborn, and K. S. Koeneman, "Predictive value of primary gleason pattern 4 in patients with gleason score 7 tumours treated with radical prostatectomy," *BJU international*, vol. 94, no. 1, pp. 42–46, 2004.

[19] W. C. Allsbrook, K. A. Mangold, M. H. Johnson, R. B. Lane, C. G. Lane, M. B. Amin, D. G. Bostwick, P. A. Humphrey, E. C. Jones, V. E. Reuter *et al.*, "Interobserver reproducibility of gleason grading of prostatic carcinoma: urologic pathologists," *Human pathology*, vol. 32, no. 1, pp. 74–80, 2001.

[20] M. W. Kattan, J. A. Eastham, A. M. Stapleton, T. M. Wheeler, and P. T. Scardino, "A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer," *JNCI: Journal of the National Cancer Institute*, vol. 90, no. 10, pp. 766–771, 1998.

[21] E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan, "Assessing the performance of prediction

models: a framework for some traditional and novel measures," *Epidemiology (Cambridge, Mass.)*, vol. 21, no. 1, p. 128, 2010.

[22] G. W. Hull, F. Rabbani, F. Abbas, T. M. Wheeler, M. W. Kattan, and P. T. Scardino, "Cancer control with radical prostatectomy alone in 1,000 consecutive patients," *The Journal of urology*, vol. 167, no. 2, pp. 528–534, 2002.

[23] M. W. Kattan, T. M. Wheeler, and P. T. Scardino, "Postoperative nomogram for disease recurrence after radical prostatectomy for prostate cancer," *Journal of Clinical Oncology*, vol. 17, no. 5, pp. 1499–1499, 1999.

[24] M. R. Cooperberg, J. M. Broering, and P. R. Carroll, "Time trends and local variation in primary treatment of localized prostate cancer," *Journal of Clinical Oncology*, vol. 28, no. 7, pp. 1117–1123, 2010.

[25] J. Ren, E. T. Sadimin, D. Wang, J. I. Epstein, D. J. Foran, and X. Qi, "Computer aided analysis of prostate histopathology images gleason grading especially for gleason score 7," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE, 2015, pp. 3013–3016.

[26] J. Ren, E. Sadimin, D. J. Foran, and X. Qi, "Computer aided analysis of prostate histopathology images to support a refined gleason grading system," in *Medical Imaging 2017: Image Processing*, vol. 10133. International Society for Optics and Photonics, 2017, p. 101331V.

[27] H. Wang, F. Xing, H. Su, A. Stromberg, and L. Yang, "Novel image markers for non-small cell lung cancer classification and survival prediction," *BMC bioinformatics*, vol. 15, no. 1, p. 310, 2014.

[28] J. Yao, S. Wang, X. Zhu, and J. Huang, "Imaging biomarker discovery for lung cancer survival prediction," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 649–657.

[29] K.-H. Yu, C. Zhang, G. J. Berry, R. B. Altman, C. Ré, D. L. Rubin, and M. Snyder, "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features," *Nature communications*, vol. 7, 2016.

[30] X. Zhu, J. Yao, and J. Huang, "Deep convolutional neural network for survival analysis with pathological images," in *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*. IEEE, 2016, pp. 544–547.

[31] X. Zhu, J. Yao, X. Luo, G. Xiao, Y. Xie, A. Gazdar, and J. Huang, "Lung cancer survival prediction from pathological images and genetic dataan integration study," in *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*. IEEE, 2016, pp. 1173–1176.

[32] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[33] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," *arXiv preprint arXiv:1702.05464*, 2017.

[34] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.

[35] X. Qi, F. Xing, D. J. Foran, and L. Yang, "Robust segmentation of overlapping cells in histopathology specimens using parallel seed detection and repulsive level set," *Biomedical Engineering, IEEE Transactions on*, vol. 59, no. 3, pp. 754–765, 2012.

[36] X. Qi, D. Wang, I. Rodero, J. Diaz-Montes, R. H. Gensure, F. Xing, H. Zhong, L. Goodell, M. Parashar, D. J. Foran *et al.*, "Content-based histopathology image retrieval using cometcloud," *BMC bioinformatics*, vol. 15, no. 1, p. 287, 2014.

[37] X. Qi, H. Kim, F. Xing, M. Parashar, D. J. Foran, and L. Yang, "The analysis of image feature robustness using cometcloud," *Journal of pathology informatics*, vol. 3, 2012.

[38] P. M. Widener, T. Kurc, W. Chen, F. Wang, L. Yang, J. Hu, V. Kumar, V. Chu, L. Cooper, J. Kong *et al.*, "High performance computing techniques for scaling image analysis workflows," in *Applied Parallel and Scientific Computing*. Springer, 2012, pp. 67–77.

[39] D. J. Foran, L. Yang, W. Chen, J. Hu, L. A. Goodell, M. Reiss, F. Wang, T. Kurc, T. Pan, A. Sharma *et al.*, "Imageminer: a software system for comparative analysis of tissue microarrays using content-based image retrieval, high-performance computing, and grid technology," *Journal of the American Medical Informatics Association*, vol. 18, no. 4, pp. 403–415, 2011.

[40] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[41] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[42] J. Ren, K. Karagoz, M. Gatza, D. J. Foran, and X. 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*, vol. 10579. International Society for Optics and Photonics, 2018, p. 1057904.

[43] J. Ren, K. Karagoz, M. L. Gatza, E. A. Singer, E. Sadimin, D. J. Foran, and X. Qi, "Recurrence analysis on prostate cancer patients with gleason score 7 using integrated histopathology whole-slide images and genomic data through deep neural networks," *Journal of Medical Imaging*, vol. 5, no. 4, p. 047501, 2018.

[44] J. Ren, J. Yang, N. Xu, and D. J. Foran, "Factorized adversarial networks for unsupervised domain adaptation," *arXiv preprint arXiv:1806.01376*, 2018.

[45] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[46] J. Ren, I. Hacihaliloglu, E. A. Singer, D. J. Foran, and X. Qi, "Adversarial domain adaptation for classification of prostate histopathology whole-slide images," *21st International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI)*, pp. 201–209, 2018.

[47] ——, "Unsupervised domain adaptation for classification of histopathology whole-slide images," *Frontiers in bioengineering and biotechnology*, vol. 7, 2019.

[48] A. Tabesh, M. Teverovskiy, H.-Y. Pang, V. P. Kumar, D. Verbel, A. Kotsianti, and O. Saidi, "Multifeature prostate cancer diagnosis and gleason grading of histological images," *IEEE transactions on medical imaging*, vol. 26, no. 10, pp. 1366–1378, 2007.

[49] K. Jafari-Khouzani and H. Soltanian-Zadeh, "Multiwavelet grading of pathological images of prostate," *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 6, pp. 697–704, 2003.

[50] P. Khurd, C. Bahlmann, P. Maday, A. Kamen, S. Gibbs-Strauss, E. M. Genega, and J. V. Frangioni, "Computer-aided gleason grading of prostate cancer histopathological images using texton forests," in *Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on.* IEEE, 2010, pp. 636–639.

[51] K. Nguyen, B. Sabata, and A. K. Jain, "Prostate cancer grading: Gland segmentation and structural features," *Pattern Recognition Letters*, vol. 33, no. 7, pp. 951–961, 2012.

[52] S. Doyle, M. Feldman, J. Tomaszewski, and A. Madabhushi, "A boosted bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies," *IEEE transactions on biomedical engineering*, vol. 59, no. 5, pp. 1205–1218, 2012.

[53] L. Gorelick, O. Veksler, M. Gaed, J. A. Gómez, M. Moussa, G. Bauman, A. Fenster, and A. D. Ward, "Prostate histopathology: Learning tissue component histograms for cancer detection and classification," *IEEE transactions on medical imaging*, vol. 32, no. 10, pp. 1804–1818, 2013.

[54] S. Kothari, J. H. Phan, R. A. Moffitt, T. H. Stokes, S. E. Hassberger, Q. Chaudry, A. N. Young, and M. D. Wang, "Automatic batch-invariant color segmentation of histological cancer images," in *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on.* IEEE, 2011, pp. 657–660.

[55] M. Macenko, M. Niethammer, J. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas, "A method for normalizing histology slides for quantitative analysis," in *Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on.* IEEE, 2009, pp. 1107–1110.

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

[57] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2015.

[58] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 11, pp. 2274–2282, 2012.

[59] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

[60] K. Nguyen, A. Sarkar, and A. K. Jain, "Structure and context in prostatic gland segmentation and classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2012, pp. 115–123.

[61] G. Lee, A. Singanamalli, H. Wang, M. D. Feldman, S. R. Master, N. N. Shih, E. Spangler, T. Rebbeck, J. E. Tomaszewski, and A. Madabhushi, "Supervised multi-view canonical correlation analysis (smvcca): integrating histologic and proteomic features for predicting recurrent prostate cancer," *IEEE transactions on medical imaging*, vol. 34, no. 1, pp. 284–297, 2015.

[62] G. Lee, R. W. Veltri, G. Zhu, S. Ali, J. I. Epstein, and A. Madabhushi, "Nuclear shape and architecture in benign fields predict biochemical recurrence in prostate cancer patients following radical prostatectomy: preliminary findings," *European urology focus*, 2016.

[63] P. Leo, G. Lee, N. N. Shih, R. Elliott, M. D. Feldman, and A. Madabhushi, "Evaluating stability of histomorphometric features across scanner and staining variations: prostate cancer diagnosis from whole slide images," *Journal of Medical Imaging*, vol. 3, no. 4, p. 047502, 2016.

[64] A. Madabhushi, S. Agner, A. Basavanhally, S. Doyle, and G. Lee, "Computer-aided prognosis: predicting patient and disease outcome via quantitative fusion of multi-scale, multi-modal data," *Computerized medical imaging and graphics*, vol. 35, no. 7-8, pp. 506–514, 2011.

[65] W. C. Allsbrook Jr, K. A. Mangold, M. H. Johnson, R. B. Lane, C. G. Lane, and J. I. Epstein, "Interobserver reproducibility of gleason grading of prostatic carcinoma: general pathologist," *Human pathology*, vol. 32, no. 1, pp. 81–88, 2001.

[66] A. Glaessgen, H. Hamberg, C.-G. Pihl, B. Sundelin, B. Nilsson, and L. Egevad, "Interobserver reproducibility of modified gleason score in radical prostatectomy specimens," *Virchows Archiv*, vol. 445, no. 1, pp. 17–21, 2004.

[67] C. Kandoth, M. D. McLellan, F. Vandin, K. Ye, B. Niu, C. Lu, M. Xie, Q. Zhang, J. F. McMichael, M. A. Wyczalkowski *et al.*, "Mutational landscape and significance across 12 major cancer types," *Nature*, vol. 502, no. 7471, p. 333, 2013.

[68] T. Makino, S. Miwa, and K. Koshida, "Impact of gleason pattern 5 on outcomes of patients with prostate cancer and iodine-125 prostate brachytherapy," *Prostate international*, vol. 4, no. 4, pp. 152–155, 2016.

[69] H. Bay, "Surf: Speeded up robust features," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[70] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

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

[72] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2. IEEE, 2005, pp. 524–531.

[73] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[74] N. Ing, Z. Ma, J. Li, H. Salemi, C. Arnold, B. S. Knudsen, and A. Gertych, "Semantic segmentation for prostate cancer grading by convolutional neural networks," in *Medical Imaging 2018: Digital Pathology*, vol. 10581. International Society for Optics and Photonics, 2018, p. 105811B.

[75] J. Xu, X. Luo, G. Wang, H. Gilmore, and A. Madabhushi, "A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images," *Neurocomputing*, vol. 191, pp. 214–223, 2016.

[76] L. Hou, K. Singh, D. Samaras, T. M. Kurc, Y. Gao, R. J. Seidman, and J. H. Saltz, "Automatic histopathology image analysis with cnns," in *Scientific Data Summit (NYSDS), 2016 New York*. IEEE, 2016, pp. 1–6.

[77] A. Cruz-Roa, A. Basavanhally, F. González, H. Gilmore, M. Feldman, S. Ganesan, N. Shih, J. Tomaszewski, and A. Madabhushi, "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks," in *SPIE medical imaging*, vol. 9041. International Society for Optics and Photonics, 2014, pp. 904 103–904 103.

[78] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in *CVPR*, 2016, pp. 2424–2433.

[79] X. Pan, L. Li, H. Yang, Z. Liu, J. Yang, L. Zhao, and Y. Fan, "Accurate segmentation of nuclei in pathological images via sparse reconstruction and deep convolutional networks," *Neurocomputing*, vol. 229, pp. 88–99, 2017.

[80] S. Naik, S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology," in *Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on*. IEEE, 2008, pp. 284–287.

[81] S. Kothari, J. H. Phan, T. H. Stokes, and M. D. Wang, "Pathology imaging informatics for quantitative analysis of whole-slide images," *Journal of the American Medical Informatics Association*, vol. 20, no. 6, pp. 1099–1108, 2013. [Online]. Available: http://dx.doi.org/10.1136/amiajnl-2012-001540

[82] V. Roullier, O. Lzoray, V.-T. Ta, and A. Elmoataz, "Multi-resolution graph-based analysis of histopathological whole slide images: Application to mitotic cell extraction and visualization," *Computerized Medical Imaging and Graphics*, vol. 35, no. 7, pp. 603 – 615, 2011, whole Slide Image Process. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0895611111000371

[83] R. Toth, N. Shih, J. Tomaszewski, M. Feldman, O. Kutter, D. Yu, J. Paulus, G. Paladini, and A. Madabhushi, "Histostitcher: An informatics software platform for reconstructing whole-mount prostate histology using the extensible imaging platform framework," *Journal of Pathology Informatics*, vol. 5, no. 1, pp. 1–9, 2014.

[84] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE, 2013, pp. 6645–6649.

[85] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 1764–1772.

[86] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[87] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[88] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue, "Modeling spatial-temporal clues in a hybrid deep learning framework for video classification," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 461–470.

[89] T. M. Therneau and P. Grambsch, "Extending the cox model," *Edited by P. Bickel, P. Diggle, S. Fienberg, K. Krickeberg*, p. 51, 2000.

[90] Y. Yang and H. Zou, "A cocktail algorithm for solving the elastic net penalized coxs regression in high dimensions," *Statistics and its Interface*, vol. 6, no. 2, pp. 167–173, 2012.

[91] J. D. Kalbfleisch and R. L. Prentice, *The statistical analysis of failure time data*. John Wiley & Sons, 2011, vol. 360.

[92] B. Moghimi-Dehkordi, A. Safaee, M. A. Pourhoseingholi, R. Fatemi, Z. Tabeie, and M. R. Zali, "Statistical comparison of survival models for analysis of cancer data," *Asian Pac J Cancer Prev*, vol. 9, no. 3, pp. 417–20, 2008.

[93] M. Cleves, *An introduction to survival analysis using Stata*. Stata Press, 2008.

[94] D. R. Cox and D. Oakes, *Analysis of survival data*. CRC Press, 1984, vol. 21.

[95] A. V. D'amico, R. Whittington, S. B. Malkowicz, D. Schultz, K. Blank, G. A. Broderick, J. E. Tomaszewski, A. A. Renshaw, I. Kaplan, C. J. Beard *et al.*, "Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer," *Jama*, vol. 280, no. 11, pp. 969–974, 1998.

[96] D. F. Gleason and G. T. Mellinger, "Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging," *The Journal of urology*, vol. 167, no. 2, pp. 953–958, 2002.

[97] H. Carter, "Active surveillance for prostate cancer: An underutilized opportunity for reducing harm," *J Natl Cancer Inst Monogr*, vol. 2012, no. 45, pp. 175–83, 2012.

[98] S. Ip, I. J. Dahabreh, M. Chung, W. W. Yu, E. M. Balk, R. C. Iovin, P. Mathew, T. Luongo, T. Dvorak, and J. Lau, "An evidence review of active surveillance in men with localized prostate cancer." *Evidence report/technology assessment*, no. 204, p. 1, 2011.

[99] J. L. Wright, C. A. Salinas, D. W. Lin, S. Kolb, J. Koopmeiners, Z. Feng, and J. L. Stanford, "Differences in prostate cancer outcomes between cases with gleason 4+ 3 and gleason 3+ 4 tumors in a population-based cohort," *The Journal of urology*, vol. 182, no. 6, p. 2702, 2009.

[100] A. Amin, A. Partin, and J. I. Epstein, "Gleason score 7 prostate cancer on needle biopsy: relation of primary pattern 3 or 4 to pathological stage and progression after radical prostatectomy," *The Journal of urology*, vol. 186, no. 4, pp. 1286–1290, 2011.

[101] M. J. Burdick, C. A. Reddy, J. Ulchaker, K. Angermeier, A. Altman, N. Chehade, A. Mahadevan, P. A. Kupelian, E. A. Klein, and J. P. Ciezki, "Comparison of biochemical relapse-free survival between primary gleason score 3 and primary gleason score 4 for biopsy gleason score 7 prostate cancer," *International Journal of Radiation Oncology* Biology* Physics*, vol. 73, no. 5, pp. 1439–1445, 2009.

[102] S. Doyle, M. Hwang, K. Shah, A. Madabhushi, M. Feldman, and J. Tomaszeweski, "Automated grading of prostate cancer using architectural and textural image features," in *Biomedical imaging: from nano to macro, 2007. ISBI 2007. 4th IEEE international symposium on*. IEEE, 2007, pp. 1284–1287.

[103] P.-W. Huang and C.-H. Lee, "Automatic classification for pathological prostate images based on fractal analysis," *IEEE transactions on medical imaging*, vol. 28, no. 7, pp. 1037–1050, 2009.

[104] D. Wang, D. J. Foran, J. Ren, H. Zhong, I. Y. Kim, and X. Qi, "Exploring automatic prostate histopathology image gleason grading via local structure modeling," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 2649–2652.

[105] C. Sotiriou and M. J. Piccart, "Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care?" *Nature Reviews Cancer*, vol. 7, no. 7, pp. 545–553, 2007.

[106] M. L. Gatza, J. E. Lucas, W. T. Barry, J. W. Kim, Q. Wang, M. D. Crawford, M. B. Datto, M. Kelley, B. Mathey-Prevot, A. Potti *et al.*, "A pathway-based classification of human breast cancer," *Proceedings of the National Academy of Sciences*, vol. 107, no. 15, pp. 6994–6999, 2010.

[107] A. H. Bild, G. Yao, J. T. Chang, Q. Wang, A. Potti, D. Chasse, M.-B. Joshi, D. Harpole, J. M. Lancaster, A. Berchuck *et al.*, "Oncogenic pathway signatures in human cancers as a guide to targeted therapies," *Nature*, vol. 439, no. 7074, pp. 353–357, 2006.

[108] M. L. Gatza, G. O. Silva, J. S. Parker, C. Fan, and C. M. Perou, "An integrated genomics approach identifies drivers of proliferation in luminal-subtype human breast cancer," *Nature genetics*, vol. 46, no. 10, pp. 1051–1059, 2014.

[109] D. G. Kleinbaum and M. Klein, *Survival analysis*. Springer, 2010, vol. 3.

[110] R. L. Grossman, A. P. Heath, V. Ferretti, H. E. Varmus, D. R. Lowy, W. A. Kibbe, and L. M. Staudt, "Toward a shared vision for cancer genomic data," *New England Journal of Medicine*, vol. 375, no. 12, pp. 1109–1112, 2016.

[111] C. E. Duchon, "Lanczos filtering in one and two dimensions," *Journal of applied meteorology*, vol. 18, no. 8, pp. 1016–1022, 1979.

[112] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[113] J. D. Hamilton, "Time series analysis," *Princeton University Press*, 1994.

[114] R. Davidson and J. G. MacKinnon., "Econometric theory and methods," *Oxford University Press*, 2004.

[115] T. A. Gerds, M. W. Kattan, M. Schumacher, and C. Yu, "Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring," *Statistics in Medicine*, vol. 32, no. 13, pp. 2173–2184, 2013.

[116] M. Wolbers, P. Blanche, M. T. Koller, J. C. Witteman, and T. A. Gerds, "Concordance for prognostic models with competing risks," *Biostatistics*, vol. 15, no. 3, pp. 526–539, 2014.

[117] B. P. Eisen MB, Spellman PT and B. D, "Cluster analysis and display of genome-wide expression patterns," *Proc Natl Acad Sci*, vol. 95, no. 25, pp. 14 863–8, 1998.

[118] Y. Zhou and J. Liu, "Ava: visual analysis of gene expression microarray data," *Bioinformatics*, vol. 19, no. 2, pp. 293–294, 2003.

[119] P. Wirapati, C. Sotiriou, S. Kunkel, P. Farmer, S. Pradervand, B. Haibe-Kains, C. Desmedt, M. Ignatiadis, T. Sengstag, F. Schütz *et al.*, "Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast

cancer subtyping and prognosis signatures," *Breast Cancer Research*, vol. 10, no. 4, p. R65, 2008.

[120] C. Fan, A. Prat, J. S. Parker, Y. Liu, L. A. Carey, M. A. Troester, and C. M. Perou, "Building prognostic models for breast cancer patients using clinical variables and hundreds of gene expression signatures," *BMC medical genomics*, vol. 4, no. 1, p. 3, 2011.

[121] A. Thorner, K. Hoadley, J. Parker, S. Winkel, R. Millikan, and C. Perou, "In vitro and in vivo analysis of b-myb in basal-like breast cancer," *Oncogene*, vol. 28, no. 5, p. 742, 2009.

[122] F. C. Herschkowitz JI, He X and P. CM., "The functional loss of the retinoblastoma tumour suppressor is a common event in basal-like and luminal b breast carcinomas," *Breast Cancer Res*, vol. 10, no. 5, p. R75, 2008.

[123] J. E. Hutti, A. D. Pfefferle, S. C. Russell, M. Sircar, C. M. Perou, and A. S. Baldwin, "Oncogenic pi3k mutations lead to nf-κb–dependent cytokine expression following growth factor deprivation," *Cancer research*, 2012.

[124] K. A. Hoadley, V. J. Weigman, C. Fan, L. R. Sawyer, X. He, M. A. Troester, C. I. Sartor, T. Rieger-House, P. S. Bernard, L. A. Carey *et al.*, "Egfr associated expression profiles vary with breast tumor subtype," *BMC genomics*, vol. 8, no. 1, p. 258, 2007.

[125] G. Bindea, B. Mlecnik, M. Tosolini, A. Kirilovsky, M. Waldner, A. C. Obenauf, H. Angell, T. Fredriksen, L. Lafontaine, A. Berger *et al.*, "Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer," *Immunity*, vol. 39, no. 4, pp. 782–795, 2013.

[126] P. Casbas-Hernandez, M. DArcy, E. Roman-Perez, H. A. Brauer, K. McNaughton, S. M. Miller, R. K. Chhetri, A. L. Oldenburg, J. M. Fleming, K. D. Amos *et al.*, "Role of hgf in epithelial–stromal cell interactions during progression from benign breast disease to ductal carcinoma in situ," *Breast Cancer Research*, vol. 15, no. 5, p. R82, 2013.

[127] M. S. Schröder, A. C. Culhane, J. Quackenbush, and B. Haibe-Kains, "survcomp: an r/bioconductor package for performance assessment and comparison of survival models," *Bioinformatics*, vol. 27, no. 22, pp. 3206–3208, 2011.

[128] Y.-S. Ha, A. Salmasi, M. Karellas, E. A. Singer, J. H. Kim, M. Han, A. W. Partin, W.-J. Kim, D. H. Lee, and I. Y. Kim, "Increased incidence of pathologically nonorgan confined prostate cancer in african-american men eligible for active surveillance," *Urology*, vol. 81, no. 4, pp. 831–836, 2013.

[129] H. B. Carter, A. W. Partin, P. C. Walsh, B. J. Trock, R. W. Veltri, W. G. Nelson, D. S. Coffey, E. A. Singer, and J. I. Epstein, "Gleason score 6 adenocarcinoma: should it be labeled as cancer?" *Journal of Clinical Oncology*, vol. 30, no. 35, p. 4294, 2012.

[130] E. A. Singer, A. Kaushal, B. Turkbey, A. Couvillon, P. A. Pinto, and H. L. Parnes, "Active surveillance for prostate cancer: past, present and future," *Current opinion in oncology*, vol. 24, no. 3, pp. 243–250, 2012.

[131] A. H. Beck, A. R. Sangoi, S. Leung, R. J. Marinelli, T. O. Nielsen, M. J. Van De Vijver, R. B. West, M. Van De Rijn, and D. Koller, "Systematic analysis of breast cancer morphology uncovers stromal features associated with survival," *Science translational medicine*, vol. 3, no. 108, pp. 108ra113–108ra113, 2011.

[132] A. C. Raldow, D. Zhang, M.-H. Chen, M. H. Braccioforte, B. J. Moran, and A. V. Damico, "Risk group and death from prostate cancer: implications for active surveillance in men with favorable intermediate-risk prostate cancer," *JAMA oncology*, vol. 1, no. 3, pp. 334–340, 2015.

[133] A. M. Khan, N. Rajpoot, D. Treanor, and D. Magee, "A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 6, pp. 1729–1738, 2014.

[134] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[135] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in *Advances in neural information processing systems*, 2013, pp. 2553–2561.

[136] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on.* IEEE, 2009, pp. 248–255.

[137] A. Gretton, A. J. Smola, J. Huang, M. Schmittfull, K. M. Borgwardt, and B. Schölkopf, "Covariate shift by kernel mean matching," 2009.

[138] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," *arXiv preprint arXiv:1612.05424*, 2016.

[139] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," *arXiv preprint arXiv:1612.07828*, 2016.

[140] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[141] J. J. Hull, "A database for handwritten text recognition research," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 16, no. 5, pp. 550–554, 1994.

[142] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *NIPS workshop on deep learning and unsupervised feature learning*, vol. 2011, no. 2, 2011, p. 5.

[143] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[144] Z. Luo, Y. Zou, J. Hoffman, and L. F. Fei-Fei, "Label efficient learning of transferable representations acrosss domains and tasks," in *Advances in Neural Information Processing Systems*, 2017, pp. 164–176.

[145] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, no. Mar, pp. 723–773, 2012.

[146] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv preprint arXiv:1412.3474*, 2014.

[147] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *International Conference on Machine Learning*, 2015, pp. 97–105.

[148] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 343–351.

[149] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *European Conference on Computer Vision*. Springer, 2016, pp. 597–613.

[150] A. Rozantsev, M. Salzmann, and P. Fua, "Beyond sharing weights for deep domain adaptation," *arXiv preprint arXiv:1603.06432*, 2016.

[151] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *Computer Vision–ECCV 2016 Workshops*. Springer, 2016, pp. 443–450.

[152] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4068–4076.

[153] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International Conference on Machine Learning*, 2015, pp. 1180–1189.

[154] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Advances in neural information processing systems*, 2016, pp. 469–477.

[155] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," *arXiv preprint arXiv:1611.02200*, 2016.

[156] B. Cheung, J. A. Livezey, A. K. Bansal, and B. A. Olshausen, "Discovering hidden factors of variation in deep networks," *arXiv preprint arXiv:1412.6583*, 2014.

[157] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2016, pp. 2172–2180.

[158] M. F. Mathieu, J. J. Zhao, J. Zhao, A. Ramesh, P. Sprechmann, and Y. LeCun, "Disentangling factors of variation in deep representation using adversarial training," in *Advances in Neural Information Processing Systems*, 2016, pp. 5040–5048.

[159] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," *arXiv preprint arXiv:1511.05644*, 2015.

[160] M. Salzmann, C. H. Ek, R. Urtasun, and T. Darrell, "Factorized orthogonal latent spaces," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 701–708.

[161] Y. Jia, M. Salzmann, and T. Darrell, "Factorized latent spaces with structured sparsity," in *Advances in Neural Information Processing Systems*, 2010, pp. 982–990.

[162] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2528–2535.

[163] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2200–2207.

[164] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, no. 2, 2017, p. 7.

[165] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[166] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," 2007.

[167] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *European conference on computer vision*. Springer, 2010, pp. 213–226.

[168] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation." in *AAAI*, vol. 6, no. 7, 2016, p. 8.

[169] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[170] M. Titford and B. Bowman, "What may the future hold for histotechnologists?" *Laboratory Medicine*, vol. 43, no. suppl_2, pp. e5–e10, 2012.

[171] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Computer graphics and applications*, vol. 21, no. 5, pp. 34–41, 2001.

[172] M. N. Gurcan, L. Boucheron, A. Can, A. Madabhushi, N. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE reviews in biomedical engineering*, vol. 2, p. 147, 2009.

[173] X. Li and K. N. Plataniotis, "A complete color normalization approach to histopathology images using color cues computed from saturation-weighted statistics," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 7, pp. 1862–1873, 2015.

[174] A. N. Basavanhally, S. Ganesan, S. Agner, J. P. Monaco, M. D. Feldman, J. E. Tomaszewski, G. Bhanot, and A. Madabhushi, "Computerized image-based detection and grading of lymphocytic infiltration in her2+ breast cancer histopathology," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 3, pp. 642–653, 2010.

[175] P. W. Hamilton, P. H. Bartels, D. Thompson, N. H. Anderson, R. Montironi, and J. M. Sloan, "Automated location of dysplastic fields in colorectal histology using image texture analysis," *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, vol. 182, no. 1, pp. 68–75, 1997.

[176] A. Ruiz, O. Sertel, M. Ujaldon, U. Catalyurek, J. Saltz, and M. Gurcan, "Pathological image analysis using the gpu: Stroma classification for neuroblastoma," in *Bioinformatics and Biomedicine, 2007. BIBM 2007. IEEE International Conference on*. IEEE, 2007, pp. 78–88.

[177] H. Qureshi, O. Sertel, N. Rajpoot, R. Wilson, and M. Gurcan, "Adaptive discriminant wavelet packet transform and local binary patterns for meningioma subtype classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2008, pp. 196–204.

[178] J. I. Epstein, M. J. Zelefsky, D. D. Sjoberg, J. B. Nelson, L. Egevad, C. Magi-Galluzzi, A. J. Vickers, A. V. Parwani, V. E. Reuter, S. W. Fine *et al.*, "A contemporary prostate cancer grading system: a validated alternative to the gleason score," *European urology*, vol. 69, no. 3, pp. 428–435, 2016.

[179] M. W. Lafarge, J. P. Pluim, K. A. Eppenhof, P. Moeskops, and M. Veta, "Domain-adversarial neural networks to address the appearance variability of histopathology images," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2017, pp. 83–91.

[180] G. Litjens, C. I. Sánchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C. Hulsbergen-Van De Kaa, P. Bult, B. Van Ginneken, and J. Van Der Laak, "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," *Scientific reports*, vol. 6, p. 26286, 2016.

[181] S. Otálora, A. Cruz-Roa, J. Arevalo, M. Atzori, A. Madabhushi, A. R. Judkins, F. González, H. Müller, and A. Depeursinge, "Combining unsupervised feature learning and riesz wavelets for histopathology image representation: application to identifying anaplastic medulloblastoma," in *MICCAI*. Springer, 2015, pp. 581–588.

[182] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A. M. Schlitter, I. Esposito, and N. Navab, "Structure-preserving color normalization and sparse stain separation for histological images," *IEEE transactions on medical imaging*, vol. 35, no. 8, pp. 1962–1971, 2016.

[183] P. Ranefall, L. Egevad, B. Nordin, and E. Bengtsson, "A new method for segmentation of colour images applied to immunohistochemically stained cell nuclei," *Analytical Cellular Pathology*, vol. 15, no. 3, pp. 145–156, 1997.

[184] C. Meurie, G. Lebrun, O. Lezoray, and A. Elmoataz, "A comparison of supervised pixels-based color image segmentation methods. application in cancerology," *WSEAS transactions on Computers*, vol. 2, no. 3, pp. 739–44, 2003.

[185] K. Z. Mao, P. Zhao, and P.-H. Tan, "Supervised learning-based cell image segmentation for p53 immunohistochemistry," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 6, pp. 1153–1163, 2006.

[186] A. Tam, J. Barker, and D. Rubin, "A method for normalizing pathology images to improve feature extraction for quantitative pathology," *Medical physics*, vol. 43, no. 1, pp. 528–537, 2016.

[187] N. Alsubaie, N. Trahearn, S. E. A. Raza, D. Snead, and N. M. Rajpoot, "Stain deconvolution using statistical analysis of multi-resolution stain colour representation," *PloS one*, vol. 12, no. 1, p. e0169875, 2017.

[188] O. J. del Toro, M. Atzori, S. Otálora, M. Andersson, K. Eurén, M. Hedlund, P. Rönnquist, and H. Müller, "Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade gleason score," in *Medical Imaging 2017: Digital Pathology*, vol. 10140. International Society for Optics and Photonics, 2017, p. 101400O.

[189] S. Sankaranarayanan, Y. Balaji, C. D. Castillo, and R. Chellappa, "Generate to adapt: Aligning domains using generative adversarial networks," in *CVPR*, 2018.

[190] A. Janowczyk, A. Basavanhally, and A. Madabhushi, "Stain normalization using sparse autoencoders (stanosa): Application to digital pathology," *Computerized Medical Imaging and Graphics*, vol. 57, pp. 50–61, 2017.

[191] M. Gadermayr, V. Appel, B. M. Klinkhammer, P. Boor, and D. Merhof, "Which way round? a study on the performance of stain-translation for segmenting arbitrarily dyed histological images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 165–173.

[192] F. G. Zanjani, S. Zinger, B. E. Bejnordi, J. A. van der Laak, and P. H. de With, "Stain normalization of histopathology images using generative adversarial networks," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 573–577.

[193] A. K. Jain, *Fundamentals of digital image processing*. Englewood Cliffs, NJ: Prentice Hall,, 1989.

[194] J. Kong, O. Sertel, H. Shimada, K. L. Boyer, J. H. Saltz, and M. N. Gurcan, "Computer-aided grading of neuroblastic differentiation: Multi-resolution and multi-classifier approach." in *ICIP (5)*, 2007, pp. 525–528.

[195] M. M. R. Krishnan, P. Shah, C. Chakraborty, and A. K. Ray, "Statistical analysis of textural features for improved classification of oral histopathological images," *Journal of medical systems*, vol. 36, no. 2, pp. 865–881, 2012.

[196] N. Papadakis, E. Provenzi, and V. Caselles, "A variational model for histogram transfer of color images," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1682–1695, 2011.

[197] J. D. Hipp, J. Y. Cheng, M. Toner, R. G. Tompkins, and U. J. Balis, "Spatially invariant vector quantization: A pattern matching algorithm for multiple classes of image subject matter including pathology," *Journal of pathology informatics*, vol. 2, 2011.

[198] A. Basavanhally and A. Madabhushi, "Em-based segmentation-driven color standardization of digitized histopathology," vol. 8676, 2013, pp. 8676 – 8676 – 12.

[199] B. E. Bejnordi, G. Litjens, N. Timofeeva, I. Otte-Höller, A. Homeyer, N. Karssemeijer, and J. A. van der Laak, "Stain specific standardization of whole-slide histopathological images," *IEEE transactions on medical imaging*, vol. 35, no. 2, pp. 404–415, 2016.

[200] M. Niethammer, D. Borland, J. S. Marron, J. Woosley, and N. E. Thomas, *Appearance normalization of histology slides*. Springer Berlin Heidelberg, 2010.

[201] M. Gavrilovic, J. C. Azar, J. Lindblad, C. Wählby, E. Bengtsson, C. Busch, and I. B. Carlbom, "Blind color decomposition of histological images." *IEEE Trans. Med. Imaging*, vol. 32, no. 6, pp. 983–994, 2013.

[202] M. D. Zarella, C. Yeoh, D. E. Breen, and F. U. Garcia, "An alternative reference space for h&e color normalization," *PloS one*, vol. 12, no. 3, p. e0174489, 2017.

[203] D. Magee, D. Treanor, D. Crellin, M. Shires, K. Smith, K. Mohee, and P. Quirke, "Colour normalisation in digital histopathology images," in *Proc Optical Tissue Image analysis in Microscopy, Histopathology and Endoscopy (MICCAI Workshop)*, vol. 100. Daniel Elson, 2009.

[204] H. Cho, S. Lim, G. Choi, and H. Min, "Neural stain-style transfer learning using gan for histopathological images," *arXiv preprint arXiv:1710.08543*, 2017.

[205] A. Bentaieb and G. Hamarneh, "Adversarial stain transfer for histopathology image analysis," *IEEE transactions on medical imaging*, vol. 37, no. 3, pp. 792–802, 2018.

[206] M. T. Shaban, C. Baur, N. Navab, and S. Albarqouni, "Staingan: Stain style transfer for digital histological images," 2018.

[207] F. G. Zanjani, S. Zinger, B. E. Bejnordi, J. A. van der Laak *et al.*, "Histopathology stain-color normalization using deep generative models," 2018.

[208] C. Wu, W. Wen, T. Afzal, Y. Zhang, Y. Chen, and H. Li, "A compact dnn: approaching googlenet-level accuracy of classification and domain adaptation," *arXiv preprint arXiv:1703.04071*, 2017.

[209] S. Herath, M. T. Harandi, and F. Porikli, "Learning an invariant hilbert space for domain adaptation." in *CVPR*, 2017, pp. 3956–3965.

[210] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo, "Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2272–2281.

[211] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray, "Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012," *International journal of cancer*, vol. 136, no. 5, 2015.

[212] P. Khurd, L. Grady, A. Kamen, S. Gibbs-Strauss, E. M. Genega, and J. V. Frangioni, "Network cycle features: Application to computer-aided gleason grading of prostate cancer histopathological images," in *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*. IEEE, 2011, pp. 1632–1636.

[213] F. Ciompi, O. Geessink, B. E. Bejnordi, G. S. de Souza, A. Baidoshvili, G. Litjens, B. van Ginneken, I. Nagtegaal, and J. van der Laak, "The importance of stain normalization in colorectal tissue classification with convolutional networks," in *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*. IEEE, 2017, pp. 160–163.

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

[215] S. Roy, A. kumar Jain, S. Lal, and J. Kini, "A study about color normalization methods for histopathology images," *Micron*, 2018.

[216] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *CVPR*, vol. 1. IEEE, 2005, pp. 539–546.

[217] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: a system for large-scale machine learning." in *OSDI*, vol. 16, 2016, pp. 265–283.

[218] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[219] Y. Liu, K. Gadepalli, M. Norouzi, G. E. Dahl, T. Kohlberger, A. Boyko, S. Venugopalan, A. Timofeev, P. Q. Nelson, G. S. Corrado *et al.*, "Detecting cancer metastases on gigapixel pathology images," *arXiv preprint arXiv:1703.02442*, 2017.

[220] A. Rakhlin, A. Shvets, V. Iglovikov, and A. A. Kalinin, "Deep convolutional neural networks for breast cancer histology image analysis," in *International Conference Image Analysis and Recognition*. Springer, 2018, pp. 737–744.

[221] K. Nazeri, A. Aminpour, and M. Ebrahimi, "Two-stage convolutional neural network for breast cancer histology image classification," in *International Conference Image Analysis and Recognition*. Springer, 2018, pp. 717–726.

[222] M. W. Fagerland, S. Lydersen, and P. Laake, "The mcnemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional," *BMC medical research methodology*, vol. 13, no. 1, p. 91, 2013.

[223] J. Diaz-Montes, M. Zou, R. Singh, S. Tao, and M. Parashar, "Data-driven workflows in multi-cloud marketplaces," in *IEEE Cloud*, 2014.

[224] J. Diaz-Montes, Y. Xie, I. Rodero *et al.*, "Federated computing for the masses - aggregating resources to tackle large-scale engineering problems," *CiSE Magazine*, vol. 16, no. 4, pp. 62–72, 2014.

[225] Z. Li and M. Parashar, "Comet: A scalable coordination space for decentralized distributed environments," in *Intl. Workshop on Hot Topics in Peer-to-Peer Systems*, 2005.

[226] G. Loy and A. Zelinsky, "Fast radial symmetry for detecting points of interest," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 8, pp. 959–973, 2003.

[227] J. Ren, Z. Li, J. Yang, N. Xu, T. Yang, and D. J. Foran, "Eigen: Ecologically-inspired genetic approach for neural network structure searching from scratch," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9059–9068.