DEEP FEATURE REPRESENTATIONS AND MULTI-INSTANCE MULTI-LABEL LEARNING OF WHOLE SLIDE BREAST HISTOPATHOLOGY IMAGES

A DISSERTATION SUBMITTED TO

THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

COMPUTER ENGINEERING

By

Caner Mercan

March 2019

DEEP FEATURE REPRESENTATIONS AND MULTI-INSTANCE MULTI-LABEL LEARNING OF WHOLE SLIDE BREAST HISTOPATHOLOGY IMAGES

By Caner Mercan

March 2019

We certify that we have read this dissertation and that in our opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

Selim Aksoy (Advisor)

Ramazan Gökberk Cinbiş

Hamdi Dibeklioğlu

Pınar Duygulu Şahin

Özgür Ulusoy

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

DEEP FEATURE REPRESENTATIONS AND MULTI-INSTANCE MULTI-LABEL LEARNING OF WHOLE SLIDE BREAST HISTOPATHOLOGY IMAGES

Caner Mercan
Ph.D. in Computer Engineering
Advisor: Selim Aksoy
March 2019

The examination of a tissue sample has traditionally involved a pathologist investigating the case under a microscope. Whole slide imaging technology has recently been utilized for the digitization of biopsy slides, replicating the microscopic examination procedure with the computer screen. This technology made it possible to scan the slides at very high resolutions, reaching up to $100,000 \times 100,000$ pixels. The advancements in the imaging technology has allowed the development of automated tools that could help reduce the workload of pathologists during the diagnostic process by performing analysis on the whole slide histopathology images.

One of the challenges of whole slide image analysis is the ambiguity of the correspondence between the diagnostically relevant regions in a slide and the slide-level diagnostic labels in the pathology forms provided by the pathologists. Another challenge is the lack of feature representation methods for the variable number of variable-sized regions of interest (ROIs) in breast histopathology images as the state-of-the-art deep convolutional networks can only operate on fixed-sized small patches which may cause structural and contextual information loss. The last and arguably the most important challenge involves the clinical significance of breast histopathology, for the misdiagnosis or the missed diagnoses of a case may lead to unnecessary surgery, radiation or hormonal therapy.

We address these challenges with the following contributions. The first contribution introduces the formulation of the whole slide breast histopathology image analysis problem as a multi-instance multi-label learning (MIMLL) task where a slide corresponds to a bag that is associated with the slide-level diagnoses provided by the pathologists, and the ROIs inside the slide correspond to the instances in the bag. The second contribution involves a novel feature representation method for the variable number of variable-sized ROIs using the activations of deep convolutional networks. Our final contribution includes a more advanced MIMLL

formulation that can simultaneously perform multi-class slide-level classification and ROI-level inference.

Through quantitative and qualitative experiments, we show that the proposed MIMLL methods are capable of learning from only slide-level information for the multi-class classification of whole slide breast histopathology images and the novel deep feature representations outperform the traditional features in fully supervised and weakly supervised settings.

Keywords: Multi-instance multi-label learning, deep convolutional features, whole slide imaging, breast histopathology, digital pathology, medical image analysis.

ÖZET

TÜM SLAYT MEME HİSTOPATOLOJİ GÖRÜNTÜLERİNİN DERİN ÖZNİTELİK GÖSTERİMLERİ VE ÇOKLU-ÖRNEK ÇOKLU-ETİKET ÖĞRENİMİ

Caner Mercan
Bilgisayar Mühendisliği, Doktora
Tez Danışmanı: Selim Aksoy
Mart 2019

Geleneksel olarak, bir doku numunesinin incelenmesi, o örneğin bir patolog tarafından mikroskop yardımıyla taranmasını içermekteydi. Tüm slayt görüntüleme teknolojisi, meme biyopsi slaytlarının bilgisayar ortamına aktarılması ile, mikroskopik inceleme sürecinin bilgisayar ekranıyla desteklenmesine olanak sağlamıştır. Bu teknoloji, slaytları çok yüksek çözünürlüklerde taramayı mümkün kılarak, bu slaytların $100,000$ piksele $100,000$ piksellik boyutlarda incelenmesine olanak tanımıştır. Görüntüleme teknolojisindeki gelişmeler, tüm slayt histopatoloji görüntüleri üzerinde analizler yaparak, teşhis sürecinde patologların iş yükünü azaltmaya yardımcı olabilecek otomatik araçların geliştirilmesini sağlamıştır.

Tüm slayt görüntü analizinin zorluklarından biri, patolog tarafından bir slayt ile ilişkilendirilmiş olan teşhisler ile slaytta bulunan bölgelerin arasındaki bağlantıların bilinmemesidir. Bunun nedeni, patologların doldurdukları patoloji formlarındaki teşhislerin slayt seviyesinde bilgi içermesidir. Diğer bir zorluk ise, tüm meme histopatolojisi görüntülerinde değişken sayıda bulunan değişken büyüklükteki ilgi bölgelerinin (İB) temsilidir. Çünkü modern evrişimli ağlar histopatoloji görüntülerindeki yapısal ve çevresel bilgiyi kodlayamayacak kadar küçüklükteki sabit boyutlu küçük pencereler üzerinde çalışmaktadır. Meme histopatolojisi görüntülerinin incelenmesindeki en büyük zorluk ise bu alanın içerdiği klinik önemden dolayıdır çünkü bir vakanın yanlış sınıflandırılması gereksiz radyasyon tedavisine, cerrahi ve hormonal tedaviye sebebiyet verebilmektedir.

Bu zorlukların ışığında, tüm slayt meme histopatolojisi görüntülerinin incelenmesini şu şekilde ele almaktayız. İlk katkı, tüm slayt meme histopatolojisinde görüntü analizi probleminin, çok-örnekli çok-etiketli bir öğrenme (ÇÖÇE) görevi

olarak tanımlanması şeklindedir. Bu bağlamda, bir torba bir slayta tekabül etmektedir ve patoloji formunda bulunan slayt etiketleriyle ilişkilendirilmektedir. Benzer şekilde, torbadaki örnekler de slayt içerisinde bulunan İB'ye tekabül etmektedir. İkinci katkı, derin evrişimli ağların özelliklerini kullanarak, değişken sayıdaki ve değişken büyüklükteki İB'nin temsili için yeni bir öznitelik gösterim yöntemini barındırmaktadır. Nihai katkımız ise, eşzamanlı olarak slayt seviyesinde çok-sınıflı sınıflandırması yapan ve İB seviyesinde tanı etiketi çıkarsayan gelişmiş bir ÇÖÇE modeli içermektedir.

Bu çalışma kapsamında geliştirilen ÇÖÇE yöntemlerinin sadece slayt düzeyinde bilgi kullanarak tüm slayt meme histopatolojisi görüntülerinin çok-sınıflı sınıflandırılması probleminde öğrenme ve genelleme yeteneğine sahip olduğunu ve ek olarak, derin öznitelik gösterimlerinin tam denetimli ve zayıf denetimli senaryolarda geleneksel öznitelik gösterimlerine kıyasla daha yüksek başarım verdiğini göstermekteyiz.

Anahtar sözcükler: Çok-örnekli çok-etiketli öğrenme, derin evrişimli öznitelik gösterimi, tüm slayt görüntüleme, meme histopatolojisi, sayısal patoloji, tıbbi görüntü analizi.

Acknowledgement

I would like to dedicate this thesis to my mother who has always believed in me more than I ever have. This would not be possible without her unconditional love, unwavering support and contagious optimism.

First and foremost, I would like to thank my advisor, Assoc. Prof. Dr. Selim Aksoy, for his patience and guidance throughout my academic career. I have learnt so much from his work ethic and I have been constantly inspired by his passion for scientific research.

I would like to thank my tracking committee members; Asst. Prof. Dr. R. Gökberk Cinbiş and Asst. Prof. Dr. Hamdi Dibeklioğlu for their time and making TTC meetings productive and enjoyable. I would also like to thank Prof. Dr. Pınar Duygulu Şahin and Prof. Dr. Özgür Ulusoy for accepting my invitation to my defence and for commenting on my thesis.

My journey has been tough at times but my friends; Fu, Huns, Fahis, Simge, Kaan and my past/present office mates have turned it into a joyful ride. I would like to specifically thank Ebru and Damla abla for always being there for me and for putting a smile on my face when I was down.

Finally, I would like to thank to the Scientific and Technological Research Council of Turkey (TÜBİTAK) for providing financial assistance throughout my PhD studies with grants 113E602 and 117E172. I would also like to thank TÜBA-GEBİP for supporting me financially on my scientific travels.

Contents

1 Introduction **1**
1.1 Motivation 2
1.2 Contribution 4

2 Related Literature **7**
2.1 Related Work for Multi-instance Multi-label Learning 7
2.2 Related Work for Deep Feature Representations 8

3 Data Set **11**
3.1 Data Set Description 11
3.2 Identification of Candidate ROIs 15

4 Multi-instance Multi-label Learning of Whole Slide Breast Histopathology Images **19**
4.1 Feature Extraction 21
4.2 Learning 22
4.3 Classification 26
4.3.1 Slide-level Classification 26
4.3.2 ROI-level Classification 27
4.4 Experimental Results 28
4.4.1 Experimental Setting 28
4.4.2 Evaluation Criteria 29
4.4.3 Slide-level Classification Results 30
4.4.4 ROI-level Classification Results 35
4.5 Discussion 40

5 Deep Feature Representations for Variable-sized ROIs in Whole Slide Images 43
5.1 Patch-level Deep Network Training . . . 45
5.1.1 Identification of Patches from ROI . . . 46
5.1.2 CNN Training on Patches . . . 47
5.2 ROI-level Deep Feature Representations . . . 47
5.2.1 ROI-level Feature Representation from Weighted Patch-level Penultimate Layer CNN Features . . . 48
5.2.2 ROI-level Feature Representation from Weighted Pixel-level Hypercolumn CNN Features . . . 49
5.3 Classification . . . 51
5.3.1 ROI-level Classification . . . 52
5.3.2 Slide-level Classification . . . 52
5.4 Experimental Results . . . 53
5.4.1 ROI-level Classification Results . . . 55
5.4.2 Slide-level Classification Results . . . 60
5.5 Discussion . . . 65

6 Joint Slide-level Multi-class Classification and ROI-level Prediction of Whole Slide Breast Histopathology Images 68
6.1 Feature Extraction . . . 70
6.1.1 Elimination of Candidate ROIs . . . 70
6.1.2 ROI-level Deep Feature Generation . . . 72
6.2 Learning . . . 74
6.2.1 Model Definition . . . 75
6.2.2 Training . . . 79
6.3 Classification . . . 84
6.3.1 Slide-level Classification . . . 84
6.3.2 ROI-level Classification . . . 85
6.4 Experimental Results . . . 85
6.4.1 Experimental Setting . . . 85
6.4.2 Evaluation Criteria . . . 88
6.4.3 Slide-level Classification Results . . . 90

6.4.4 ROI-level Classification Results 92
6.5 Discussion 95

7 Conclusion **98**

List of Figures

1.1 The framework depicting the steps of the analysis and the classification of a whole slide breast histopathology image 4

3.1 Viewing behavior of eight different pathologists on a whole slide image 14
3.2 ROI detection from the viewport logs of a pathologist 16
3.3 An example slide with ROI annotations and diagnostic labels involving individual pathologists and their consensus labels 18

4.1 Feature extraction process for an example ROI 22
4.2 Different learning scenarios in the context of whole slide breast histopathology 24
4.3 Whole slide ROI-level classification example for a case with ADH 37
4.4 Whole slide ROI-level classification example for a case with DCIS 38

5.1 Patch selection process shown on an example ROI 46
5.2 The deep feature generation process 48
5.3 The feature vector from the convolutional channels 52
5.4 The ROI-level feature representation steps 53
5.5 The architecture of the VGG16 network with batch normalization 54
5.6 Patch-level classification outputs 61
5.7 Example ROI proposals and consensus ROIs 63
5.8 ROI-level classification outputs on example slides 66

6.1 The overview of the proposed HCRF based MIMLL approach . . 70
6.2 The discovery of candidate ROIs in whole slide images 73
6.3 The overview of the ROI-level feature generator network 74

6.4 The graphical view of the proposed model 78
6.5 The visual representation of the Contrastive Divergence algorithm 81
6.6 The converged values of the parameters of the model 89
6.7 The ROI-level predictions of the MimlHCRF model on example whole slide images 93

List of Tables

3.1 Distribution of diagnostic classes among the 240 slides 13
3.2 Hierarchical mapping of the original 14 classes of diagnoses to subsets of 5 and 4 classes 15

4.1 Summary of the features for each candidate ROI 21
4.2 Summary statistics for the number of candidate ROIs extracted from the viewing logs 27
4.3 5-class slide-level average precision classification average precision results of the experiments with a particular pathologist's data . . 30
4.4 5-class slide-level classification results of the experiments with the union of three pathologists' data 31
4.5 14-class slide-level classification average precision results of the experiments with a particular pathologist's data 33
4.6 14-class slide-level classification results of the experiments with the union of three pathologists' data 34
4.7 Confusion matrix for ROI-level classification 36
4.8 Class-specific statistics on the performance of ROI-level classification 36
4.9 Kappa scores for 5-class classification 41
4.10 Kappa scores for 14-class classification 42

5.1 The class distribution of the slides and the ROIs 55
5.2 The comparison of ROI-level classification performance 57
5.3 Confusion matrix of Penultimate-Feat-Weighted for ROI-level classification 58
5.4 Class-specific statistics for Penultimate-Feat-Weighted features . . 58

5.5 Confusion matrix of Hypercolumn-Feat-Weighted for ROI-level classification 59
5.6 Class-specific statistics for Hypercolumn-Feat-Weighted features . 59
5.7 The slide-level classification performance comparison 62
5.8 The confusion matrix of the slide-level classification results with Penultimate-Feat-Weighted features 64
5.9 The confusion matrix of the slide-level classification results with Hypercolumn-Feat-Weighted features 64

6.1 The distribution of the most severe diagnostic categories in the data set 86
6.2 Statistics of class combinations in the training and test data . . . 86
6.3 Comparison of the slide-level multi-class classification results . . . 94
6.4 The confusion matrix of the ROI-level classification results 95
6.5 Class-specific statistics on the performance of ROI-level classification 95

Chapter 1

Introduction

The diagnosis for cancer is traditionally made through a microscopic examination of a tissue sample by highly-trained pathologists. In recent years, the field of pathology has seen a huge paradigm shift from glass slides to whole slide imaging with the advancements in digital imaging technology. Whole slide imaging is an automated technology that has allowed glass slides to be scanned at high resolutions to produce very large digital slides. Digitization of full biopsy slides using the whole slide imaging technology has provided new opportunities for understanding the diagnostic process of pathologists and developing more accurate computer aided diagnosis systems. The alarmingly increasing number of cancer patients has also necessitated automated systems that could aid the pathologists and reduce their workload during their screening of the slides. Histopathological image analysis has shown great potential in supporting the diagnostic process for cancer by providing objective and repeatable measures for characterizing the tissue samples to reduce the observer variations in the diagnoses [1].

The typical approach for computing these measures is to use statistical classifiers that are built by employing supervised learning algorithms on data sets that involve carefully selected regions of interest (ROIs) with diagnostic labels assigned by pathologists. Furthermore, performance evaluation of these methods has also been limited to the use of manually chosen image areas that correspond

to isolated tissue structures with no ambiguity regarding their diagnoses. Such studies that are built around these restricted training and test settings do not necessarily reflect the complexity of the decision process encountered in routine histopathological examinations. In this thesis, our main focus is to address the complexity of the entire diagnostic process with the development of state-of-the-art algorithms that are weakly trained on data available only at slide-level to provide diagnostic predictions to an unseen slide as well as to its parts.

1.1 Motivation

Breast histopathology is one particular example with a continuum of histologic features that have different clinical significance. For example, proliferative changes such as usual ductal hyperplasia (UDH) are considered benign, and patients diagnosed with UDH do not undergo any additional procedures [2]. On the other hand, major clinical treatment thresholds exist between atypical ductal hyperplasia (ADH) and ductal carcinoma in situ (DCIS) that carry different risks of progressing into malignant invasive carcinoma [3]. In particular, when a biopsy that actually has ADH is overinterpreted as DCIS, a person may undergo unnecessary surgery, radiation, and hormonal therapy [4]. These problems are even more important considering that breast cancer is the most prevalent type of cancer among women and the second most common overall with over 2 million cases observed only in 2018.[1]

The varying degrees of clinical significance of cases in breast histopathology and the improved imaging tools that output high resolution digital breast biopsy images have accelerated the development of image analysis systems to aid the pathologists in their interpretation of the slides. However, these approaches suffer from several drawbacks.

First, the generalizability of the image analysis algorithms with accuracies reported for the simplified setting of benign versus malignant cases is not applicable

[1] `https://www.wcrf.org/dietandcancer/cancer-trends/breast-cancer-statistics`

for the finer-grained categorization problem involving a greater clinical significance. The screening of a slide involves a pathologist interpreting the slide thoroughly. Based on her/his observations, she/he finalizes the diagnostic procedure by filling out the associated pathology form with one or more of the diagnostic labels. Contrary to the traditional algorithms that perform binary categorization of benign vs. malignant cases, a slide can be categorized under one or more of the several diagnostic labels with varying clinical significance. Therefore, the categorization of a whole slide image turns into a much more challenging and rewarding problem when multi-class and multi-label analysis of slides are considered.

Second, the algorithms trained on manually selected patches extracted from the ROIs in the whole slide images do not reflect the real-world procedure of whole slide histopathology interpretation and diagnosis. During her/his interpretation of the slides, the pathologist tries to spot the diagnostically relevant regions inside a slide to investigate such regions in more detail. A slide can include several such regions with varying degrees of diagnostic relevance. After her/his interpretation of the slide, the pathologist fills out the pathology form of the slide based on her/his discoveries in the diagnostically relevant regions. Therefore, there is no manually selected patch or patch-level information available; as a matter of fact, there is no correspondence between the regions inside a slide and the diagnostic labels provided only at slide-level in the real-world clinical setting. The investigation of the ambiguity between slide parts and the slide-level diagnostic information involves a more realistic scenario for the whole slide image analysis task.

Finally, learning a representation of a whole slide image in its entirety is impossible due to its inherently large size. One of the more commonly used approaches is to represent a slide as a bag of its fixed-sized small patches. Even though this kind of representation is commonly employed due to the restrictions of the state-of-the-art feature generator architectures, such formulation does not consider the contextual or the structural information of the ROIs in the slide. A slide can contain variable number of variable-sized ROIs that play pivotal role in the overall diagnoses of the slide. Hence, there is a need for the development of a set of algorithms that are capable of capturing the properties of variable-sized

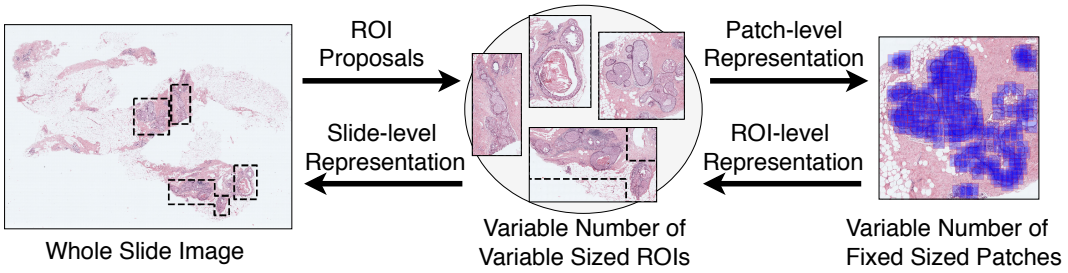


Figure 1.1: The framework depicting the steps of the analysis and the classification of a whole slide breast histopathology image. A whole slide image is represented by a variable number of variable-sized ROIs, each ROI can be characterized as a bag of potentially informative small fixed-sized patches. The patch-level information builds ROI-level knowledge which in turn is used to make a prediction for the whole slide image.

ROIs by fully exploiting the state-of-the-art patch-level deep feature generator architectures to improve the diagnostic procedure of the whole slide images.

1.2 Contribution

The field of whole slide breast histopathology analysis has seen very limited investigation in scenarios that involve patient cases with varying degrees of clinical significance and machine learning methodologies that reflect the complexity of the decision process that the pathologists go through during histopathological examinations. In this thesis, we tackle the multi-class classification task of whole slide breast histopathology images considering varying degrees of clinical significance and the challenging procedure inherent to the routine whole slide histopathology screening by proposing state-of-the-art multi-instance multi-label learning algorithms involving novel deep feature representations. Our whole slide breast histopathology analysis and classification frameworks follow the process presented in Figure 1.1. In this regard, this thesis has the following three main contributions.

Our first contribution is the introduction of the first multi-instance multi-label learning approach for the multi-class classification of the whole slide breast histopathology images [5,6]. We used certain actions defined on the viewing logs of

the pathologists involving screen coordinates of the inspected slide area coupled with time stamps which were recorded when they were interpreting the slides. The procedure allowed us to locate regions in the slides that were potentially diagnostically relevant. Our first contribution involves the formulation of a multi-instance multi-label learning scenario in which a bag is a whole slide image and the instances of the bag are the regions outlined by the actions of the pathologists. We compute color, texture and nuclear features to represent the instances (ROIs) in a bag (slide). Each bag is also associated with slide-level labels that the pathologists fill out in the pathology form. Multi-instance multi-label learning algorithms are trained on these bags using the associated slide-level labels to perform extensive evaluations on different experimental settings to form a baseline for the weakly supervised learning of whole slide breast histopathology images. This method is described in greater detail in Chapter 4.

For our second contribution, we devise a feature generator network to represent the variable-sized ROIs in whole slide images [7]. The slides contain variable number of ROIs, and each of those ROIs can play a vital role in the diagnosis of the slide. Learning feature representations of these regions using state-of-the-art convolutional networks are neither trivial nor straightforward as the sizes of such regions can differ greatly and they all have arbitrary shapes. We introduce a novel approach that can generate deep feature representations for ROIs regardless of their shapes or sizes. The method involves the aggregation of feature vectors of fixed-sized small patches, which are automatically extracted from the potentially informative areas inside the region, that are weighted by class specific patch predictions using the properties of a single fine-tuned convolutional network. We train classifiers on the proposed feature representations to perform multi-class classification and demonstrate that the proposed approach outperforms the existing ones through quantitative and qualitative experiments. This method is presented in Chapter 5.

Our third and the final contribution involves a multi-instance multi-label learning framework that jointly models complex relations and associations of ROIs and their latent labels, and coherence as well as correlations between latent region labels and slide labels in a weakly supervised learning scenario. The deep feature

generator network is also incorporated to represent the ROIs within the slides. The model is capable of inferring region labels from slide-level information by considering the individual properties of the ROIs, their spatial distribution inside the slide, and the coherence of the predicted labels of the regions with the slide labels. We investigate the performance of deep convolutional feature representations compared to the traditional hand-crafted features based on color, texture and nuclear architecture in the context of multi-instance multi-label learning. More importantly, we demonstrate the power of the proposed model, which simultaneously performs multi-class slide-level classification and infers the diagnostic labels of the ROIs, by comparing its slide-level classification performance with the previous best efforts. Note that the previous works were not capable of making predictions at ROI-level; therefore, we could only compare the slide-level performance. The methodology of the proposed approach is discussed comprehensively in Chapter 6.

The remainder of this thesis is organized as follows. The related literature involving weakly supervised learning of whole slide histopathology images with an emphasis on breast histopathology and involving deep feature representations of whole slide images is presented in Chapter 2. The data set description and data preprocessing steps are detailed in Chapter 3. We present the baseline multi-instance multi-label learning algorithms for the multi-class classification of whole slide images in Chapter 4. Chapter 5 describes the deep feature generator network for learning representations of ROIs. A more sophisticated and flexible formulation of multi-instance multi-label learning for the simultaneous multi-class classification of whole slide images and prediction of ROIs is deeply discussed in Chapter 6. Finally, the summary of all studies in this thesis and the future work are given in Chapter 7.

Chapter 2

Related Literature

2.1 Related Work for Multi-instance Multi-label Learning

The use of multi-instance and multi-label learning algorithms has been quite rare in the field of histopathological image analysis. Dundar et al. [8] presented one of the first applications of multi-instance learning for breast histopathology by designing a large margin classifier for binary discrimination of benign cases from actionable (ADH+DCIS) ones by using whole slides with manually identified ROIs. For the same binary classification task, Xu et al. [9] used boosting-based multi-instance learning to predict cases as benign or cancer. In similar fashion, square tissue patches were used as instances for multi-instance classification of tissue images as healthy or cancer [10]. The same group also incorporated relational learning to multi-instance learning for the binary classification task [11]. In different subdomains of histopathology for whole slide image analysis, Cosatto et al. [12] performed binary classification of gastric cancer in a multi-instance learning framework. In addition, joint patch-level segmentation and slide-level binary classification of histopathology images of colon cancer was explored by Xu et al. [13]. In one of the earliest works involving multi-label learning, multi-label

support vector machines were studied for multi-class classification of colon cancer by Xu et al. [14]. More recently, Han et al. [15] proposed a multi-label classification method for breast histopathology involving a deep network optimizing the intra-class and inter-class distances that aimed to learn feature representations from ROIs. In the context of multi-instance learning involving histopathology image analysis, conditional random fields have seen limited coverage. Binary classification of whole slide breast histopathology images was performed in a setting that involved conditional random fields by Zanjani et al. [16]. Another work addressed the mitosis detection task in whole slide breast histopathology images exploiting the spatial correlations of neighboring patches using a neural conditional random field [17].

Most of the related work involving histopathological image analysis consider either multi-instance learning or multi-label learning scenarios. The works involving multi-instance learning only focus on the less clinically significant task of binary classification of whole slide images as cancer versus non-cancer. The first application of both multi-instance and multi-label learning for the slide-level multi-class classification of breast histopathology was introduced by Mercan et al. [5] and the same group presented an extension of the work involving nuclear architecture features on top of color and texture features with support for ROI-level classification [6].

2.2 Related Work for Deep Feature Representations

Convolutional neural networks are the foundations of many state-of-the-art methods for computer vision tasks, including the recent methods in histopathology image analysis [18–20]. However, their limitation in input size dictating that it should be of specific size and shape delayed their application in the medical imaging domain. This has been a very relevant problem for the histopathological image analysis as ROIs in whole slide images tend to have arbitrary shapes

and typically very large sizes. Previous works generally avoided this problem by training convolutional networks on fixed-sized cropped patches sampled from the slides [21–23]. However, this resulted in a severe loss of contextual information in the ROIs. Some other works resized the ROIs to the required input size of the convolutional networks by downsampling which resulted in a significant loss of structural information in the ROIs [24, 25]. The loss of contextual and structural information becomes more significant when the challenging multi-class classification setting is concerned, as opposed to the more restricted binary benign vs. cancer setup.

Other works in the histopathological image analysis using deep networks focused on binary (benign vs. malignant tissue) classification problem. In one of the first works in the field, Cruz et al. [26] showed that the classification of regions with invasive cancer using a convolutional architecture outperformed the existing classification methods trained on hand-crafted features. Alexnet [27], previously one of the most popular CNN architectures, was trained on the manually selected patches of histopathology images to classify the non-overlapping grid patches in a whole slide image [21], for which the slide-level predictions were performed by simple fusion rules on the patch-level predictions. Simultaneously detecting the magnification of the patches was also studied as a side task to the binary class classification problem previously [22, 23].

Classifying a tissue as one of the many subcategories of breast cancer is clinically more significant than the binary classification of cases as benign vs. malignant. There have been multiple works addressing the multi-class classification of breast histopathology images involving convolutional networks. Four-class classification of breast histopathology images was performed using a convolutional network on manually selected patches, and a slide-level prediction was made by combining patch-level classification outputs with specific fusion rules [28]. An ensemble fusion framework involving a logistic regression classifier was adopted in another work to solve the more general four-class classification problem at slide-level by exploiting the patch-level CNN probabilities [29]. In another work by Hou et al. [30], the patch-level probabilities from the network were combined to create a class frequency histogram which was then used to represent the slide

that the patches were extracted from. For the multi-class classification task, a classifier was trained on the class frequency histograms of the slides using the associated slide labels to analyze breast histopathology images.

Aside from using the convolutional networks directly for classification, its representational power was also investigated in several works. Convolutional features were extracted with a constraint that emphasizes inter-class differences while keeping intra-class differences small in a binary classification setting [31]. Xu et al. [32] trained stacked sparse auto-encoders [33] on nuclei of breast histopathology images to learn high-level features that encoded contextual information. Zheng et al. [34] encoded nucleus-centered patches by a set of individual auto-encoder units which were then stacked to learn slide-level feature representations. More recently, the representational capabilities of deep convolutional networks were combined with the traditional hand-crafted features for structure prediction in medical images for the segmentation task [35].

Most of the existing works operate only on individual fixed-sized small patches, while others use simple fusion-based approaches such as majority voting and aggregation of patch-level probabilities to perform slide-level classification. None of the existing works consider that the ROIs play pivotal role in the diagnostic process of a whole slide image and the feature representations should be based on structural and contextual information at ROI-level.

Chapter 3

Data Set

The data used in this study were collected in the scope of an NIH-sponsored project titled "Digital Pathology, Accuracy, Viewing Behavior and Image Characterization (digiPATH)" [1] that aims to evaluate the accuracy of pathologists' interpretation of digital images vs. glass slides.

3.1 Data Set Description

We used 240 haematoxylin and eosin (H&E) stained slides of breast biopsies that were selected from two registries that were associated with the Breast Cancer Surveillance Consortium [36]. Each slide belonged to an independent case from a different patient where a random stratified method was used to include cases that covered the full range of diagnostic categories from benign to invasive cancer. The class composition is given in Table 3.1. The cases with atypical ductal hyperplasia and ductal carcinoma in situ were intentionally oversampled to gain statistical precision in the estimation of interpretive concordance for these diagnoses [4].

The selected slides were scanned at 40× magnification, resulting in an average

[1] `https://homes.cs.washington.edu/~shapiro/digipath.html`

image size of 100,000 × 64,000 pixels. The cases were randomly assigned to one of four test sets, each including 60 cases with the same class frequency distribution, by using stratified sampling based on age, breast density, original reference diagnosis, and experts' difficulty rating of the case [36]. A total of 87 pathologists were recruited to evaluate the slides, and one of the four test sets was randomly assigned to each pathologist. Thus, each slide has, on average, independent interpretations from 22 pathologists. The data collection also involved tracking pathologists' actions while they were interpreting the slides using a web-based software tool that allowed seamless multi-resolution browsing of image data. The tracking software recorded the screen coordinates and mouse events at a frequency of four entries per second. At the end of the viewing session, each participant was also asked to provide a diagnosis by selecting one or more of the 14 classes on a pathology form to indicate what she/he had seen during her/his screening of the slide. Data for an example slide are illustrated in Figure 3.1. Throughout this study, we mostly use a more general set of five classes as well as another set of four classes with their mappings from the original diagnoses shown in Table 3.2.

In addition, three experienced pathologists who are internationally recognized for research and education on diagnostic breast pathology evaluated every slide both independently and in consensus meetings where the result of the consensus meeting was accepted as the reference diagnoses for each slide. The difficulty of the classification problem studied here can be observed from the evaluation presented in [37] where the individual pathologists' concordance rates compared with the consensus-derived reference diagnosis was 82% for the union of non-proliferative and proliferative changes, 43% for ADH, 79% for DCIS, and 93% for invasive carcinoma. In our experiments, we only used the individual viewing logs and the diagnostic classifications from the three experienced pathologists for slide-level evaluation, because they were the only ones who evaluated all of the 240 slides. These pathologists' data also contained a bounding box around an example region that corresponded to the most representative and supporting ROI for the most severe diagnosis that was observed during their examination of that slide during consensus meetings. We refer to the annotated regions in a slide as

Table 3.1: Distribution of diagnostic classes among the 240 slides. 14-class distribution includes all labels in the pathology form whereas 5-class and 4-class distributions involve only the most severe diagnostic label in the slide.

(a) 14-class distribution

Class	# slides
Non-proliferative changes only	7
Fibroadenoma	16
Intraductal papilloma w/o atypia	11
Usual ductal hyperplasia	65
Columnar cell hyperplasia	89
Sclerosing adenosis	18
Complex sclerosing lesion	9
Flat epithelial atypia	37
Atypical ductal hyperplasia	69
Intraductal papilloma w/ atypia	15
Atypical lobular hyperplasia	18
Ductal carcinoma in situ	89
Lobular carcinoma in situ	7
Invasive carcinoma	22

(b) 5-class distribution

Class	# slides
Non-proliferative changes only	13
Proliferative changes	63
Atypical ductal hyperplasia	66
Ductal carcinoma in situ	76
Invasive carcinoma	22

(c) 4-class distribution

Class	# slides
Benign without atypia	56
Atypical ductal hyperplasia	83
Ductal carcinoma in situ	79
Invasive carcinoma	22

Figure 3.1: Viewing behavior of eight different pathologists on a whole slide image with a size of 74896×75568 pixels. The time spent by each pathologist on different image areas is illustrated using the heat map given above the images. The unmarked regions represent unviewed areas, and overlays from dark gray to red and yellow represent increasing cumulative viewing times. The diagnostic labels assigned by each pathologist to this image are also shown.

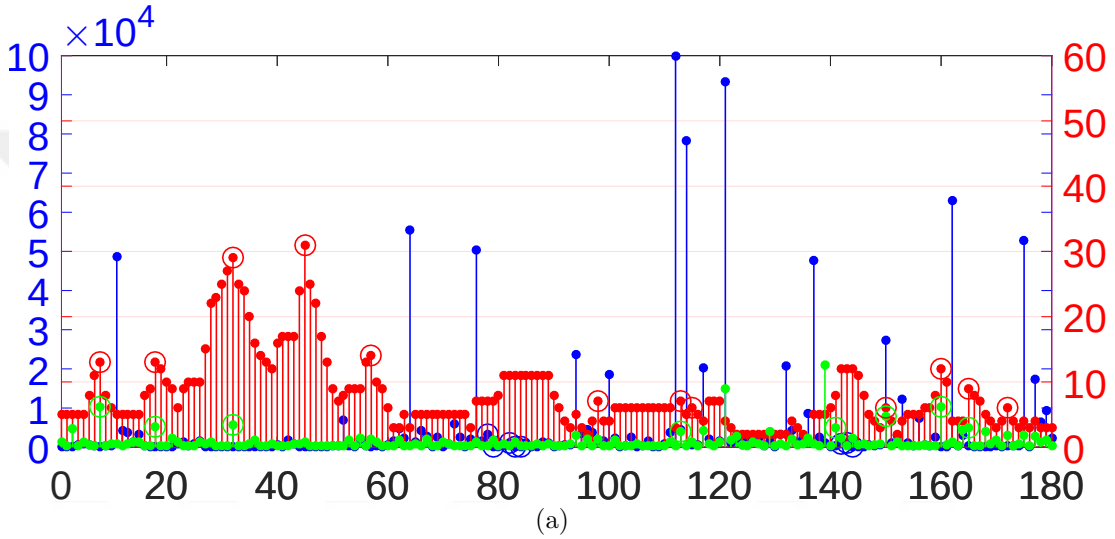
Table 3.2: Hierarchical mapping of original 14 diagnoses to subsets of 5 and 4 classes. The mappings were designed by experienced pathologists [36]. The subset with 4-classes involved both lobular and ductal malignancies and the focus in the subset with 5-classes was to study ductal malignancies, so when only lobular carcinoma in situ or atypical lobular hyperplasia was present in a slide, it was put to the non-proliferative category.

Diagnosis	5-class mapping	4-class mapping
Non-proliferative changes only	NonProliferative	Benign
Fibroadenoma	NonProliferative	Benign
Intraductal papilloma w/o atypia	Proliferative	Benign
Usual ductal hyperplasia	Proliferative	Benign
Columnar cell hyperplasia	Proliferative	Benign
Sclerosing adenosis	Proliferative	Benign
Radial scar complex sclerosing lesion	Proliferative	Benign
Flat epithelial atypia	Proliferative	Atypical
Atypical ductal hyperplasia	Atypical	Atypical
Intraductal papilloma with atypia	Atypical	Atypical
Atypical lobular hyperplasia	NonProliferative	Atypical
Ductal carcinoma in situ	DCIS	DCIS
Lobular carcinoma in situ	NonProliferative	DCIS
Invasive carcinoma	Invasive	Invasive

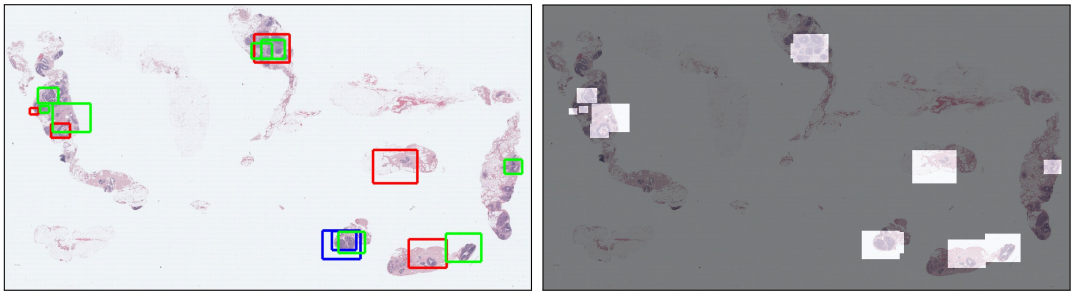
the consensus ROIs and to the associated most severe reference diagnosis as the consensus label of the slide. There are 437 such ROIs. Each consensus ROIs is assumed to have the same label as the slide-level consensus diagnosis.

3.2 Identification of Candidate ROIs

Following the observation that different pathologists have different interpretive viewing behaviors [38, 39], the following three actions were defined: *zoom peak* is an entry that corresponds to an image area where the pathologist investigated closer by zooming in, and is defined as a local maximum in the zoom level; *slow panning* corresponds to image areas that are visited in consecutive viewports where the displacement (measured as the difference between the center pixels of two viewports) is small while the zoom level is constant; *fixation* corresponds to an area that is viewed for more than 2 seconds. The union of all viewports that belonged to one of these actions was selected as the set of candidate ROIs. Figure



(a)

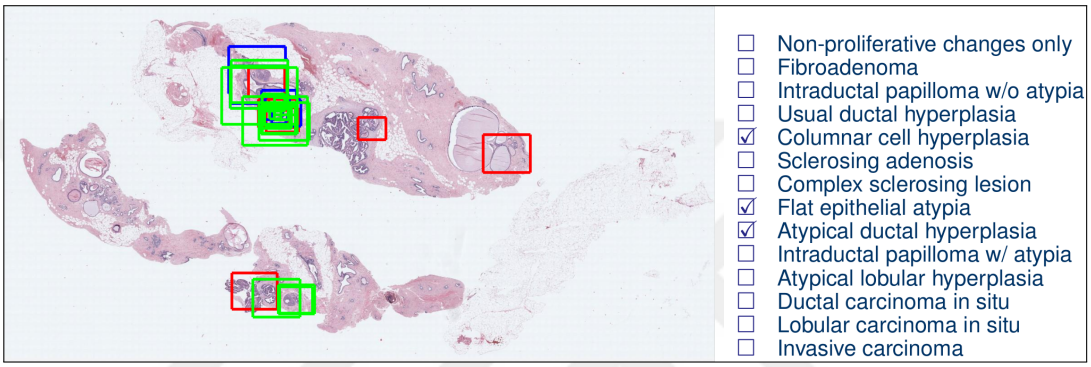


(b) (c)

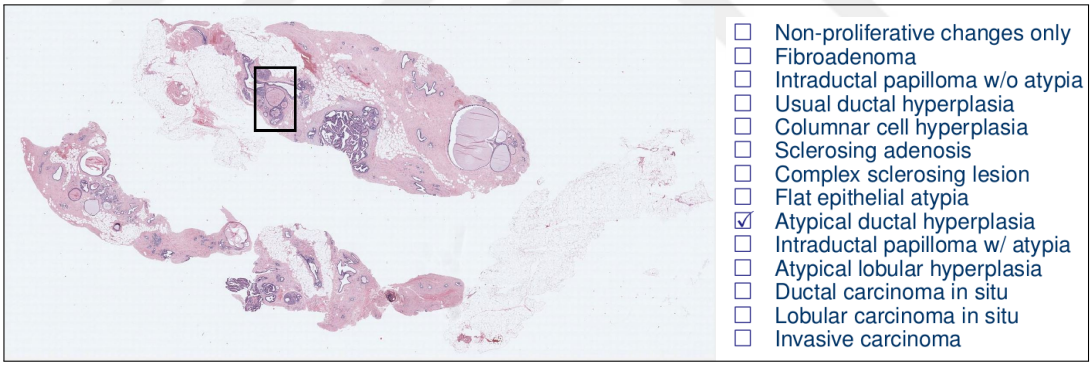
Figure 3.2: ROI detection from the viewport logs. (a) Viewport log of a particular pathologist. The x-axis shows the log entry. The red, blue, and green bars represent the zoom level, displacement, and duration, respectively. (b) The rectangular regions visible on the pathologist's screen during the selected actions are drawn on the actual image. A *zoom peak* is a red circle in (a) and a red rectangle in (b), a *slow panning* is a blue circle in (a) and a blue rectangle in (b), a *fixation* is a green circle in (a) and a green rectangle in (b). (c) Candidate ROIs resulting from the union of the selected actions.

3.2 illustrates the selection process for an example slide.

In summary, we used the three experienced pathologists' viewing logs, their individual assessments, and the consensus diagnoses for the four sets of 60 slides described above in experimental settings so that the training and test slides always belonged to different patients. An example slide belonging to a patient is presented in Figure 3.3. The candidate ROIs defined from the three actions and the consensus ROIs from the consensus meetings of the pathologists were shown in the same slide. In addition, the combined label set of the three pathologists as well as the reference diagnoses of the slide from the consensus meetings of the pathologists were shown in the figure.



(a)



(b)

Figure 3.3: An example slide with ROI annotations and diagnostic labels, in two different ways. (a) The first set included the union of candidate ROIs computed from the three actions (zoom peak, slow panning, fixation) from the viewing logs of the three pathologists. The candidate ROIs were associated with the combined label sets provided by the three pathologists during their individual screenings of the slide. (b) The second set included the consensus ROI/s from the results of the consensus meetings held by the three pathologists. The consensus ROI/s were associated with the most severe reference label, i.e. consensus label, provided by the three pathologists in their consensus meetings.

Chapter 4

Multi-instance Multi-label Learning of Whole Slide Breast Histopathology Images

This chapter introduces a framework that exploits the pathologists' viewing records of whole slide images and integrates them with the pathology reports for *weakly supervised learning* of fine-grained classifiers. Whole slide scanners that create high-resolution images with sizes reaching to $100{,}000 \times 100{,}000$ pixels by digitizing the entire glass slides at $40\times$ magnification have enabled the whole diagnostic process to be completed in digital format. Earlier studies that used whole slide images have focused on efficiency issues where classifiers previously trained on labeled ROI were run on large images by using multi-resolution [40] or multi-field-of-view [41] frameworks. However, two new challenges emerging from the use of whole slide images still need to be solved. The first challenge is the uncertainty regarding the correspondence between the image areas and the diagnostic labels assigned by the pathologists at the slide level. In clinical practice, the diagnosis is typically recorded for the entire slide, and the local tissue characteristics that grabbed the attention of the pathologist and led to that particular diagnosis are not known. The second challenge is the need for simultaneous detection and classification of diagnostically relevant areas in whole slides;

large images often contain multiple regions with different levels of significance for malignancy, and it is not known a priori which local cues should be classified together. Both the former challenge that is related to the learning problem and the latter challenge that corresponds to the localization problem necessitate the development of new algorithms for whole slide histopathology.

The framework uses multi-instance multi-label learning to build both slide-level and ROI-level classifiers for breast histopathology. Multi-instance learning (MIL) differs from traditional learning scenarios by use of the concept of bags, where each training bag contains several instances of positive and negative examples for the associated bag-level class label. A positive bag is assumed to contain at least one positive instance, whereas all instances in a negative bag are treated as negative examples, but the labels of the individual instances are not known during training. Multi-label learning (MLL) involves the scenarios in which each training example is associated with more than one label, as it can be possible to describe a sample in multiple ways. Multi-instance multi-label learning (MIMLL) corresponds to the combined case where each training sample is represented by a bag of multiple instances, and the bag is assigned multiple class labels. Most of the related studies in the literature consider only either the MIL [8–10, 12] or the MLL [14, 15] scenario. In this study, we present experimental results on the categorization of breast histopathology images into 5 and 14 classes in a weakly supervised setting.

The main contributions of this study are twofold. First, we study the MIMLL scenario in the context of whole slide image analysis. In our scenario, a bag corresponds to a digitized breast biopsy slide, the instances correspond to candidate ROIs in the slide, and the class labels correspond to the diagnoses associated with the slide. The candidate ROIs are identified by using a rule-based analysis of recorded actions of pathologists while they were interpreting the slides. The class labels are extracted from the forms that the pathologists filled out according to what they saw during their interpretation of the image. The second contribution is an extensive evaluation of the performances of four MIMLL algorithms on multi-class prediction of both the slide-level (bag-level) and the ROI-level

Table 4.1: Summary of the features for each candidate ROI. Nuclear architecture features were derived from the Voronoi diagram (VD), Delaunay triangulation (DT), minimum spanning tree (MST), and nearest neighbor (NN) statistics of nuclei centroids. The number of features is given for each type.

Type	Description
Lab (192)	64-bin histogram of the CIE-L channel 64-bin histogram of the CIE-a channel 64-bin histogram of the CIE-b channel
LBP (128)	64-bin histogram of the LBP codes of the H channel 64-bin histogram of the LBP codes of the E channel
VD (13)	Total area of polygons Polygon area: mean, std dev, min/max ratio, disorder Polygon perimeter: mean, std dev, min/max ratio, disorder Polygon chord length: mean, std dev, min/max ratio, disorder
DT (8)	Triangle side length: mean, std dev, min/max ratio, disorder Triangle area: mean, std dev, min/max ratio, disorder
MST (4)	Edge length: mean, std dev, min/max ratio, disorder
NN (25)	Nuclear density Distance to 3, 5, 7 nearest nuclei: mean, std dev, disorder # of nuclei in 10, 20, 30, 40, 50 μm radius: mean, std dev, disorder

(instance-level) labels for novel slides and simultaneous localization and classification of diagnostically relevant regions in whole slide images. The quantitative evaluation uses multiple performance criteria computed for classification scenarios involving 5 and 14 diagnostic classes and different combinations of viewing records from multiple pathologists. This study marks the first work that uses the MIMLL framework for learning and classification tasks involving such a comprehensive distribution of challenging diagnostic classes in histopathological image analysis.

4.1 Feature Extraction

The weakly supervised learning scenario employed in this study used candidate ROIs that were extracted from the pathologists' viewing logs as potentially informative areas that may be important for the diagnosis of the whole slide. These candidate ROIs were identified among the viewports that were sampled from the

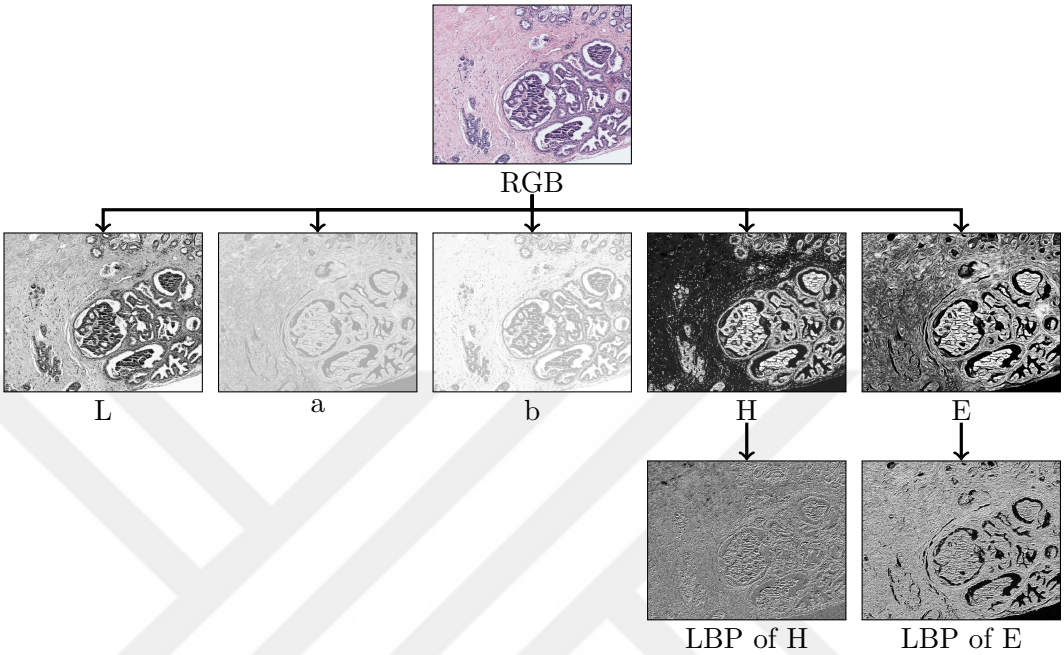


Figure 4.1: Feature extraction process for an example ROI. Contrast enhancement was performed for better visualization.

viewing session of the pathologists and were represented by the coordinates of the image area viewed on the screen, the zoom level, and the time stamp as described in Section 3.2.

The feature representation for each candidate ROI used the color histogram computed for each channel in the CIE-Lab space and texture histograms of local binary patterns computed for the haematoxylin and eosin channels estimated using a color deconvolution procedure [42]. Figure 4.1 shows the feature extraction process for an example candidate ROI. In addition, nuclear architectural features [41] computed from the nucleus detection results of [43] were also used. Table 4.1 provides the details of the resulting 370-dimensional feature vector.

4.2 Learning

The granularity of the annotations available in the training data determines the amount of supervision that can be incorporated into the learning process. Among the most popular weakly labeled learning scenarios, multi-instance learning (MIL)

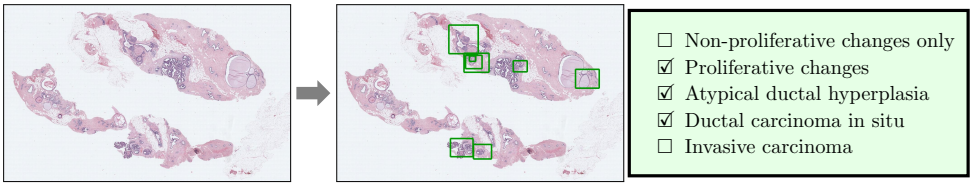
involves samples where each sample is represented by a collection (bag) of instances with a single label for the collection, and multi-label learning (MLL) uses samples where each sample has a single instance that is described by more than one label. In this section, we define the multi-instance multi-label learning (MIMLL) framework that contains both cases. Figure 4.2 illustrates the different learning scenarios in the context of whole slide imaging.

Let $\{(\mathcal{X}_m, \mathcal{Y}_m)\}_{m=1}^{M}$ be a data set with M samples where each sample consists of a bag and an associated set of labels. The bag $\mathcal{X}_m$ contains a set of instances $\{\mathbf{x}_{mn}\}_{n=1}^{n_m}$ where $\mathbf{x}_{mn} \in \mathbb{R}^d$ is the feature vector of the $n'th$ instance, and n_m is the total number of instances in that bag. The label set $\mathcal{Y}_m$ is composed of class labels $\{y_{ml}\}_{l=1}^{l_m}$ where $y_{ml} \in \{c_1, c_2, \ldots, c_L\}$ is one of L possible labels, and l_m is the total number of labels in that set. The traditional supervised learning problem is a special case of MIMLL where each sample has a single instance and a single label, resulting in the data set $\{(\mathbf{x}_m, y_m)\}_{m=1}^{M}$. MIL is also a special case of MIMLL where each bag has only one label, resulting in the data set $\{(\mathcal{X}_m, y_m)\}_{m=1}^{M}$. MLL is another special case where the single instance corresponding to a sample is associated with a set of labels, resulting in the data set $\{(\mathbf{x}_m, \mathcal{Y}_m)\}_{m=1}^{M}$.

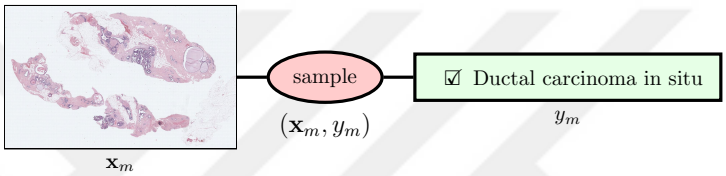
In the following, we summarize four different approaches adapted from the machine learning literature for the solution of the MIMLL problem in this study.

1. MimlSvmMi: A possible solution is to approximate the MIMLL problem as a multi-instance single label learning problem. Given an MIMLL data set with M samples, we can create a new MIL data set with $M \times \left(\sum_{m=1}^{M} l_m\right)$ samples where a sample $(\mathcal{X}_m, \mathcal{Y}_m)$ in the former is decomposed into a set of l_m bags as $\{(\mathcal{X}_m, y_{ml})\}_{l=1}^{l_m}$ in the latter by assuming that the labels are independent from each other. The resulting MIL problem is further reduced into a traditional supervised learning problem by assuming that each instance in a bag has an equal and independent contribution to the label of that bag, and is solved by using the MiSvm algorithm [44].

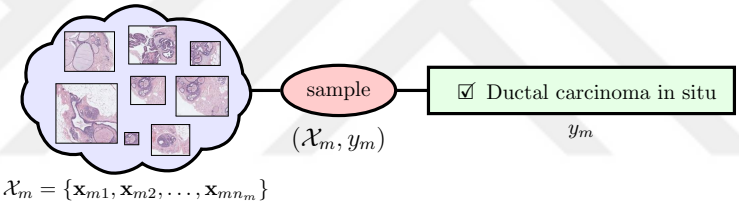
2. MimlSvm: An alternative is to decompose the MIMLL problem into a single-instance multi-label learning problem by embedding the bags in a



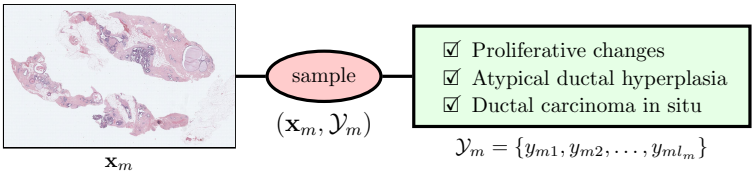
(a) Input to a learning algorithm



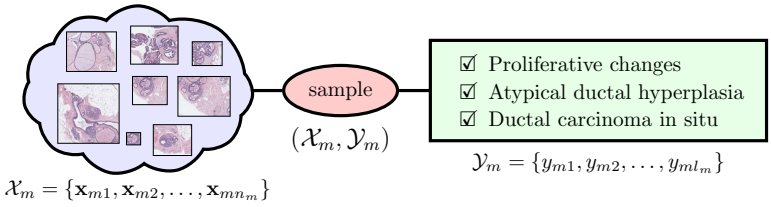
(b) Traditional supervised learning scenario



(c) Multi-instance learning (MIL) scenario



(d) Multi-label learning (MLL) scenario



(e) Multi-instance multi-label learning (MIMLL) scenario

Figure 4.2: Different learning scenarios in the context of whole slide breast histopathology. The input to a learning algorithm is the set of candidate ROIs obtained from the viewing logs of the pathologists and the diagnostic labels assigned to the whole slide. Different learning algorithms use these samples in different ways during training. The notation is defined in the text. The 5-class setting is shown, but we also use 14-class labels in the experiments.

new vector space. First, the bags are collected into a set $\{\mathcal{X}_m\}_{m=1}^{M}$, and the set is clustered using the k-medoids algorithm [45]. During clustering, the distance between two bags $\mathcal{X}_i = \{\mathbf{x}_{in}\}_{n=1}^{n_i}$ and $\mathcal{X}_j = \{\mathbf{x}_{jn}\}_{n=1}^{n_j}$ is computed by using the Hausdorff distance [46]:

$$h(\mathcal{X}_i, \mathcal{X}_j) = \max\left\{ \max_{\mathbf{x}_i \in \mathcal{X}_i} \min_{\mathbf{x}_j \in \mathcal{X}_j} \|\mathbf{x}_i - \mathbf{x}_j\|, \max_{\mathbf{x}_j \in \mathcal{X}_j} \min_{\mathbf{x}_i \in \mathcal{X}_i} \|\mathbf{x}_j - \mathbf{x}_i\| \right\}. \quad (4.1)$$

Then, the set of bags is partitioned into K clusters, each of which is represented by its medoid $\mathcal{M}_k, k = 1, \ldots, K$, the object in each cluster whose average dissimilarity to all other objects in the cluster is minimal. Finally, the embedding of a bag $\mathcal{X}_m$ into a K-dimensional space is performed by computing a vector $\mathbf{z}_m \in \mathbb{R}^K$ whose components are the Hausdorff distances between the bag and the medoids as $\mathbf{z}_m = (h(\mathcal{X}_m, \mathcal{M}_1), h(\mathcal{X}_m, \mathcal{M}_2), \ldots, h(\mathcal{X}_m, \mathcal{M}_K))$ [47]. The resulting MLL problem for the data set $\{(\mathbf{z}_m, \mathcal{Y}_m)\}_{m=1}^{M}$ is further reduced into a binary supervised learning problem for each class by using all samples that have a particular label in their label set as positive examples and the rest of the samples as negative examples for that label, and is solved using the MLSVM algorithm [48].

3. MIMLNN: Similar to MIMLSVM, the initial MIMLL problem is decomposed into an MLL problem by vector space embedding. This algorithm differs in the last step in which the resulting MLL problem is solved by using a linear classifier whose weights are estimated by minimizing a sum-of-squares error function [49].

4. M^{3}MIML: This method is motivated by the observation that useful information between instances and labels could be lost during the transformation of the MIMLL problem into an MIL (the first method) or an MLL (the second and third methods) problem [50]. The M^{3}MIML algorithm uses a linear model for each label where the output for a bag for a particular label is the maximum discriminant value among all instances of that bag under the model for that label. During training, the margin of a sample for a label is defined as this maximum over all instances, the margin of the sample for the multi-label classifier is defined as the minimum margin over all labels,

and a quadratic programming problem is solved to estimate the parameters of the linear model by maximizing the margin of the whole training set that is defined as the minimum of all samples' margins.

Each algorithm described in this section was used to learn a multi-class classifier for which each training sample was a whole slide that was modeled as a bag of candidate ROIs ($\mathcal{X}_m$), each ROI being represented by a feature vector ($\mathbf{x}_{mn}$), and a set of labels that were assigned to that slide ($\mathcal{Y}_m$). The resulting classifiers were used to predict labels for a new slide as described in the following section.

4.3 Classification

Classification was performed both at the slide level and at the ROI level. Both schemes involved the same training procedures described in Section 4.2 using the MIMLL algorithms.

4.3.1 Slide-level Classification

Given a bag of ROIs, $\mathcal{X}$, for an unknown whole slide image, a classifier trained as in Section 4.2 assigned a set of labels, $\mathcal{Y}'$, for that image. In the experiments, the bag $\mathcal{X}$ corresponded to the set of candidate ROIs extracted from the pathologists' viewing logs as described in Section 3.2. If no logs were available at test time, an ROI detector for identifying and localizing diagnostically relevant areas as described in [38] and [39] would be used. Automated ROI detection is an open problem because visual saliency (that can be modeled by well-known algorithms in computer vision) does not always correlate well with diagnostic saliency [51]. New solutions for ROI detection such as [52] can directly be incorporated in our framework to identify the candidate ROIs.

Table 4.2: Summary statistics (average ± standard deviation) for the number of candidate ROIs extracted from the viewing logs. The statistics are given for subsets of the slides for individual diagnostic classes based on the consensus labels (Non-proliferative changes only (NP), Proliferative changes (P), Atypical ductal hyperplasia (ADH), Ductal carcinoma in situ (DCIS), Invasive carcinoma (INV)) as well as the whole data set. *All* corresponds to the union of three pathologists' ROIs for a particular slide.

Class	E1	E2	E3	All
NP	13.692±14.255	22.615±21.635	6.692±7.157	43.000±32.964
P	26.507±18.734	58.285±46.989	25.333±22.514	110.127±74.218
ADH	26.500±18.355	49.227±42.374	17.863±16.469	93.590±63.545
DCIS	16.000±13.126	31.618±27.813	9.513±9.196	57.131±40.820
INV	24.409±9.163	25.9545±14.025	6.045±6.440	56.409±21.317
Whole	22.291±16.760	42.454±38.822	15.491±16.997	80.237±61.046

4.3.2 ROI-level Classification

In many previously published studies, classification at the ROI level involves manually selected regions of interest. However, this cannot be easily generalized to the analysis of whole slide images that involve many local areas that can have very different diagnostic relevance and structural ambiguities which may lead to disagreements among pathologists regarding their class assignments.

In this study, a sliding window approach for classification at the ROI level was employed. Each whole slide image was processed within sliding windows of 3600×3600 pixels with an overlap of 2400 pixels along both horizontal and vertical dimensions. The sizes of the sliding windows were determined based on the empirical observations in [38] and [39]. Each window was considered as an instance whose feature vector $\mathbf{x}$ was obtained as in Section 4.1. The classifiers learned in the previous section then assigned a set of labels $\mathcal{Y}'$ and a confidence score for each class for each window independently. Because of the overlap, each final unique classification unit corresponded to a window of 1200×1200 pixels, whose classification scores for each class were obtained by taking the per-class maximum of the scores of all sliding windows that overlap with this 1200×1200 pixel region.

4.4 Experimental Results

4.4.1 Experimental Setting

The parameters for the algorithms were set based on trials on a small part of the data, based on suggestions made in the cited papers. Three of the four algorithms (MimlSvmMi, MimlSvm, and M^{3}Miml) used support vector machines (SVM) as the base classifier. The scale parameter in the Gaussian kernel was set to 0.2 for all three methods. The number of clusters (K) in MimlSvm and MimlNN was set to 20% and 40%, respectively, of the number of training samples (bags), and the regularization parameter in the least-squares problem in MimlNN was set to 1.

The three experienced pathologists whose viewing logs were used in the experiments are denoted as *E1*, *E2*, and *E3*. For each one, the set of candidate ROIs for each slide was obtained as in Section 3.2, and the feature vector for each ROI was extracted as in Section 4.1 to form the bag of instances for that slide. The multi-label set was formed by using the labels assigned to the slide by that expert. Overall, a slide contained, on average, 1.77 ± 0.66 labels for five classes and 2.66 ± 1.29 labels for 14 classes when the label sets assigned by all experts were combined. Each slide also had a single consensus label that was assigned jointly by the three pathologists.

Table 4.2 summarizes the ROI statistics in the data set. There are some significant differences in the screening patterns of the pathologists; some spend more time on a slide and investigate a larger number of ROIs, whereas some make faster decisions by looking at a few key areas. It is important to note that the slides with consensus diagnoses of proliferative changes and atypical ductal hyperplasia attracted significantly longer views resulting in more ROIs for all pathologists.

4.4.2 Evaluation Criteria

Quantitative evaluation was performed by comparing the labels predicted for a slide by an algorithm to the labels assigned by the pathologists. The four test sets described in Section 3.1 were used in a four-fold cross-validation setup where the training and test samples (slides) came from different patients. Given the test set that consisted of N samples $\{(\mathcal{X}_n, \mathcal{Y}_n)\}_{n=1}^{N}$ where $\mathcal{Y}_n$ was the set of reference labels for the *n'th* sample, let $f(\mathcal{X}_n)$ be a function that returns the set of labels predicted by an algorithm for $\mathcal{X}_n$ and $r(\mathcal{X}_n, y)$ be the rank of the label y among $f(\mathcal{X}_n)$ when the labels are sorted in descending order of confidence in prediction (the label with the highest confidence has a rank of 1). We computed the following five criteria that are commonly used in multi-label classification:

- $hammingLoss(f) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L}|f(\mathcal{X}_n) \triangle \mathcal{Y}_n|$, where $\triangle$ is the symmetric distance between two sets. It is the fraction of wrong labels (i.e., false positives or false negatives) to the total number of labels.

- $rankingLoss(f) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{|\mathcal{Y}_n||\overline{\mathcal{Y}_n}|}|\{(y_1, y_2)|r(\mathcal{X}_n, y_1) \geq r(\mathcal{X}_n, y_2),\ (y_1, y_2) \in \mathcal{Y}_n \times \overline{\mathcal{Y}_n}\}|$, where $\overline{\mathcal{Y}_n}$ denotes the complement of the set $\mathcal{Y}_n$. It is the fraction of label pairs where a wrong label has a smaller (better) rank than a reference label.

- $one\text{-}error(f) = \frac{1}{N}\sum_{n=1}^{N}\mathbb{1}\left[\arg\min_{y\in\{c_1,c_2,\ldots,c_L\}}\ r(\mathcal{X}_n, y) \notin \mathcal{Y}_n\right]$, where $\mathbb{1}$ is an indicator function that is 1 when its argument is true, and 0 otherwise. It counts the number of samples for which the top-ranked label is not among the reference labels.

- $coverage(f) = \frac{1}{N}\sum_{n=1}^{N}\max_{y\in\mathcal{Y}_n} r(\mathcal{X}_n, y) - 1$. It is defined as how far one needs to go down the list of predicted labels to cover all reference labels.

- $averagePrecision(f) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{|\mathcal{Y}_n|}\sum_{y\in\mathcal{Y}_n}|\{y'|\ r(\mathcal{X}_n, y') \leq r(\mathcal{X}_n, y),\ y' \in \mathcal{Y}_n\}|/r(\mathcal{X}_n, y)$. It is the average fraction of correctly predicted labels that have a smaller (or equal) rank than a reference label.

To illustrate the evaluation criteria, consider a classification problem involving

Table 4.3: 5-class slide-level classification average precision results of the experiments when a particular pathologist's data (candidate ROIs and class labels) were used for training (rows) and each individual pathologist's data were used for testing (columns). The best result for each column is marked in bold.

		E1	*E2*	*E3*
E1	MimlSvmMi	0.7094 ± 0.0600	0.6253 ± 0.0584	0.6326 ± 0.0153
	MimlSvm	0.7757 ± 0.0419	0.6577 ± 0.0453	0.6901 ± 0.0060
	MimlNN	$\mathbf{0.7823 \pm 0.0332}$	0.6813 ± 0.0323	0.7113 ± 0.0215
	M^{3}Miml	0.7420 ± 0.0476	0.5922 ± 0.0450	0.6702 ± 0.0162
E2	MimlSvmMi	0.6524 ± 0.0174	0.5956 ± 0.0197	0.5908 ± 0.0243
	MimlSvm	0.7664 ± 0.0381	$\mathbf{0.6905 \pm 0.0383}$	0.6932 ± 0.0168
	MimlNN	0.7565 ± 0.0296	0.6737 ± 0.0279	0.7117 ± 0.0396
	M^{3}Miml	0.7471 ± 0.0345	0.6073 ± 0.0604	0.6993 ± 0.0245
E3	MimlSvmMi	0.6406 ± 0.0521	0.5599 ± 0.0278	0.5971 ± 0.0400
	MimlSvm	0.7570 ± 0.0239	0.6569 ± 0.0363	$\mathbf{0.7322 \pm 0.0083}$
	MimlNN	0.7657 ± 0.0199	0.6705 ± 0.0175	0.7233 ± 0.0135
	M^{3}Miml	0.7449 ± 0.0505	0.6102 ± 0.0357	0.6745 ± 0.0119

the labels $\{A, B, C, D, E\}$. Let a bag $\mathcal{X}$ have the reference labels $\mathcal{Y} = \{A, B, D\}$, and an algorithm predict $f(\mathcal{X}) = \{B, E, A\}$ in descending order of confidence. Hamming loss is $2/5 = 0.4$ (because D is a false negative and E is a false positive), ranking loss is $2/6 = 0.33$ (because (A, E) and (D, E) are wrongly ranked pairs), one-error is 0, coverage is 3 (assuming that D comes after A in the order of confidence), and average precision is $(2/3 + 1 + 3/4)/3 = 0.806$. Smaller values for the first four criteria and a larger value for the last one indicate better performance.

4.4.3 Slide-level Classification Results

The quantitative results given in this section show the average and standard deviation of the corresponding criteria computed using cross-validation. For each fold, the number of training samples, M, is 180, and the number of independent test samples, N, is 60.

Table 4.4: 5-class slide-level classification results of the experiments when the union of three pathologists' data (candidate ROIs and class labels) were used for training (rows). Test labels consisted of the union of pathologists' individual labels as well as their consensus labels in two separate experiments. The evaluation criteria are: Hamming loss (HL), ranking loss (RL), one-error (OE), coverage (COV), and average precision (AP). The best result for each setting is marked in bold.

	Test data: *E1* ∪ *E2* ∪ *E3*				
	HL	RL	OE	COV	AP
MimlSvmMi	0.3367 ± 0.0122	0.3361 ± 0.0197	0.4125 ± 0.0551	2.0542 ± 0.0798	0.7058 ± 0.0190
MimlSvm	0.2675 ± 0.0164	0.2045 ± 0.0222	0.2958 ± 0.0438	1.6917 ± 0.0967	0.7790 ± 0.0228
MimlNN	$\mathbf{0.2375 \pm 0.0189}$	$\mathbf{0.1771 \pm 0.0194}$	$\mathbf{0.2708 \pm 0.0498}$	$\mathbf{1.5583 \pm 0.0096}$	$\mathbf{0.8068 \pm 0.0262}$
M^{3}Miml	0.2842 ± 0.0152	0.2611 ± 0.0488	0.3250 ± 0.0518	1.9500 ± 0.1790	0.7301 ± 0.0374
	Test data: *Consensus*				
	HL	RL	OE	COV	AP
MimlSvmMi	0.3042 ± 0.0117	0.3528 ± 0.0096	0.5167 ± 0.0593	1.7333 ± 0.1667	0.6518 ± 0.0250
MimlSvm	0.2783 ± 0.0197	0.2295 ± 0.0351	0.4250 ± 0.1221	1.3958 ± 0.0774	0.7161 ± 0.0624
MimlNN	$\mathbf{0.2567 \pm 0.0255}$	$\mathbf{0.2049 \pm 0.0421}$	$\mathbf{0.4125 \pm 0.1181}$	$\mathbf{1.2792 \pm 0.1031}$	$\mathbf{0.7377 \pm 0.0577}$
M^{3}Miml	0.2650 ± 0.0244	0.2792 ± 0.0812	0.4583 ± 0.1251	1.5833 ± 0.2289	0.6802 ± 0.0864

4.4.3.1 5-class Classification Results

Two experiments were performed to study scenarios involving different pathologists. The goal of the first experiment was to see how well a classifier built by using only a particular pathologist's viewing records (candidate ROIs and class labels) on the training slides could predict the class labels assigned by individual pathologists to the test slides. Table 4.3 shows the average precision values for the experiments repeated using the data for each of the three pathologists separately. The results showed that MimlNN and MimlSvm performed the best, followed by M^{3}Miml, with MimlSvmMi having the worst performance. An expected result (illustrated by the columns of Table 4.3) was that the classifier that performed the best on the test data labeled by a particular pathologist was the one that was learned from the training data of the same pathologist (different slides but labeled by the same person). Among the three pathologists, the first one had the largest average number of labels assigned to the slides (1.55 labels compared to 1.20 for the second and 1.26 for the third), that probably boosted the average precision values of the classifiers on the test data of the first pathologist.

The goal of the second experiment was to evaluate the effect of diversifying the training data, where the instance set for each training slide corresponded to the union of all candidate ROIs of the three pathologists (the last row of Table 4.2), and the label set was formed as the union of all three pathologists' labels for that slide. As test labels, we used the union of three pathologists' labels as one setting, and the consensus diagnosis as another setting for each test slide. Table 4.4 shows the resulting performance statistics. The highest average precision of 0.8068 was obtained when the test labels were formed from the union of all pathologists' data. The more difficult setting that tried to predict the consensus label for each test slide resulted in an average precision of 0.7377 with MimlNN as the classifier. (The consensus label-based evaluation is harsher on wrong classifications than multi-label evaluation when at least some of the labels are predicted correctly.)

Table 4.5: 14-class slide-level classification average precision results of the experiments when a particular pathologist's data (candidate ROIs and class labels) were used for training (rows) and each individual pathologist's data were used for testing (columns). The best result for each column is marked in bold.

		E1	*E2*	*E3*
E1	MimlSvmMi	0.5154 ± 0.0399	0.4443 ± 0.0774	0.4509 ± 0.0460
	MimlSvm	0.6485 ± 0.0124	0.4950 ± 0.0370	0.5051 ± 0.0406
	MimlNN	$\mathbf{0.6787 \pm 0.0354}$	0.5243 ± 0.0258	$\mathbf{0.5534 \pm 0.0425}$
	M^{3}Miml	0.6019 ± 0.0237	0.3828 ± 0.0429	0.4342 ± 0.0573
E2	MimlSvmMi	0.4864 ± 0.0683	0.4470 ± 0.0447	0.4415 ± 0.0210
	MimlSvm	0.4953 ± 0.0637	0.5524 ± 0.0708	0.5035 ± 0.0402
	MimlNN	0.5671 ± 0.0503	$\mathbf{0.5724 \pm 0.0451}$	0.5412 ± 0.0269
	M^{3}Miml	0.5363 ± 0.0685	0.5139 ± 0.0555	0.5011 ± 0.0692
E3	MimlSvmMi	0.3988 ± 0.0587	0.3914 ± 0.0335	0.4196 ± 0.0353
	MimlSvm	0.5455 ± 0.0339	0.5262 ± 0.0523	0.5387 ± 0.0289
	MimlNN	0.5891 ± 0.0362	0.5194 ± 0.0335	0.5448 ± 0.0264
	M^{3}Miml	0.5837 ± 0.0757	0.5242 ± 0.0347	0.5414 ± 0.0171

4.4.3.2 14-class classification results

We used the same experimental setup in Section 4.4.3.1 for 14-class classification. Table 4.5 shows the average precision values. MimlNN and MimlSvm, that both formulated the MIMLL problem by embedding the bags into a new vector space and reducing it to an MLL problem, consistently outperformed both MimlSvmMi that transformed the MIMLL problem into an MIL problem by assuming independence of labels, and M^{3}Miml that used a more complex model that was more sensitive to the amount of training data. Due to similar reasons as in the previous section, the scores when the first pathologist's test data were used were higher than the scores on the test data of the second and third pathologists. Also similar to the 5-class classification results, a particular pathologist's test data were classified the best by the classifier learned from the same pathologist's training data with the exception of the third one's test data which were classified the best when the training data of the first one were used. However, the best classification performance of the third pathologist's classifier, 0.5448, was very close to the first one's classifier's best classification score of 0.5534. These experiments once again confirmed the difficulty of whole slide learning and classification by using slide-level information compared to the same by using manually selected,

Table 4.6: 14-class slide-level classification results of the experiments when the union of three pathologists' data (candidate ROIs and class labels) were used for training (rows). Test labels consisted of the union of pathologists' individual labels as well as their consensus labels in two separate experiments. The evaluation criteria are: Hamming loss (HL), ranking loss (RL), one-error (OE), coverage (COV), and average precision (AP). The best result for each setting is marked in bold.

	Test data: *E1* ∪ *E2* ∪ *E3*				
	HL	RL	OE	COV	AP
MimlSvmMi	0.2054 ± 0.0171	0.3138 ± 0.0406	0.5500 ± 0.0793	6.4750 ± 0.1917	0.5425 ± 0.0353
MimlSvm	0.1646 ± 0.0134	0.1912 ± 0.0239	0.3542 ± 0.0786	5.5208 ± 0.7504	0.6432 ± 0.0293
MimlNN	$\mathbf{0.1604 \pm 0.0103}$	$\mathbf{0.1591 \pm 0.0120}$	$\mathbf{0.3125 \pm 0.0685}$	$\mathbf{5.0042 \pm 0.4267}$	$\mathbf{0.6917 \pm 0.0307}$
M^{3}Miml	0.1732 ± 0.0041	0.2297 ± 0.0209	0.3833 ± 0.0491	6.1667 ± 0.4587	0.5661 ± 0.0324
	Test data: *Consensus*				
	HL	RL	OE	COV	AP
MimlSvmMi	0.1792 ± 0.0186	0.3093 ± 0.0326	0.6625 ± 0.0658	5.4083 ± 0.6198	0.4864 ± 0.0323
MimlSvm	0.1592 ± 0.0119	0.2181 ± 0.0211	0.5750 ± 0.1206	4.8417 ± 0.8742	0.5281 ± 0.0402
MimlNN	$\mathbf{0.1557 \pm 0.0080}$	$\mathbf{0.1843 \pm 0.0089}$	$\mathbf{0.5208 \pm 0.1109}$	$\mathbf{4.2333 \pm 0.5418}$	$\mathbf{0.5855 \pm 0.0456}$
M^{3}Miml	0.1565 ± 0.0044	0.2618 ± 0.0165	0.6125 ± 0.1117	5.4833 ± 0.5307	0.4568 ± 0.0507

well-defined regions as commonly studied in the literature.

The second set of experiments followed the same procedure as in Section 4.4.3.1 as well. Table 4.6 presents the quantitative results. In agreement with the 5-class classification results, the best performance was achieved when the union of all pathologists' data were used for both training and testing, but with a drop in average precision from 0.8068 to 0.6917 for the more challenging 14-class setting. We would like to note that it was not straightforward to compare the 5-class and 14-class performances with respect to all evaluation criteria, as the number of labels in the respective test sets could often be different, and some performance criteria (e.g., coverage) were known to be more sensitive to the number of labels than others.

4.4.4 ROI-level Classification Results

We followed the sliding window approach described in Section 4.3.2 to obtain confidence scores for all classes at each 1200×1200 pixel window of a whole slide image. The best performing classifier of the previous section, MimlNN, was selected for training with the union of all candidate ROIs from the three pathologists and with the slide-level consensus labels. We used only the 5-class setting, since the consensus reference data used for performance evaluation at the ROI level had only 5-class information.

As mentioned in Section 3.1, ADH and DCIS cases were oversampled during data set construction [4]. This made automatic learning of the minority classes NP and INV difficult even though they are relatively easier for humans. Therefore, we employed an upsampling approach for these two classes where a new bag was formed by sampling with replacement from the instances of a randomly selected bag until the number of training samples increased by twofold. The resulting set was used for weakly-labeled training of a multi-class classifier from slide-level information for ROI-level classification.

Since only the diagnostic label of the consensus ROI was known for each slide,

Table 4.7: Confusion matrix for ROI-level classification.

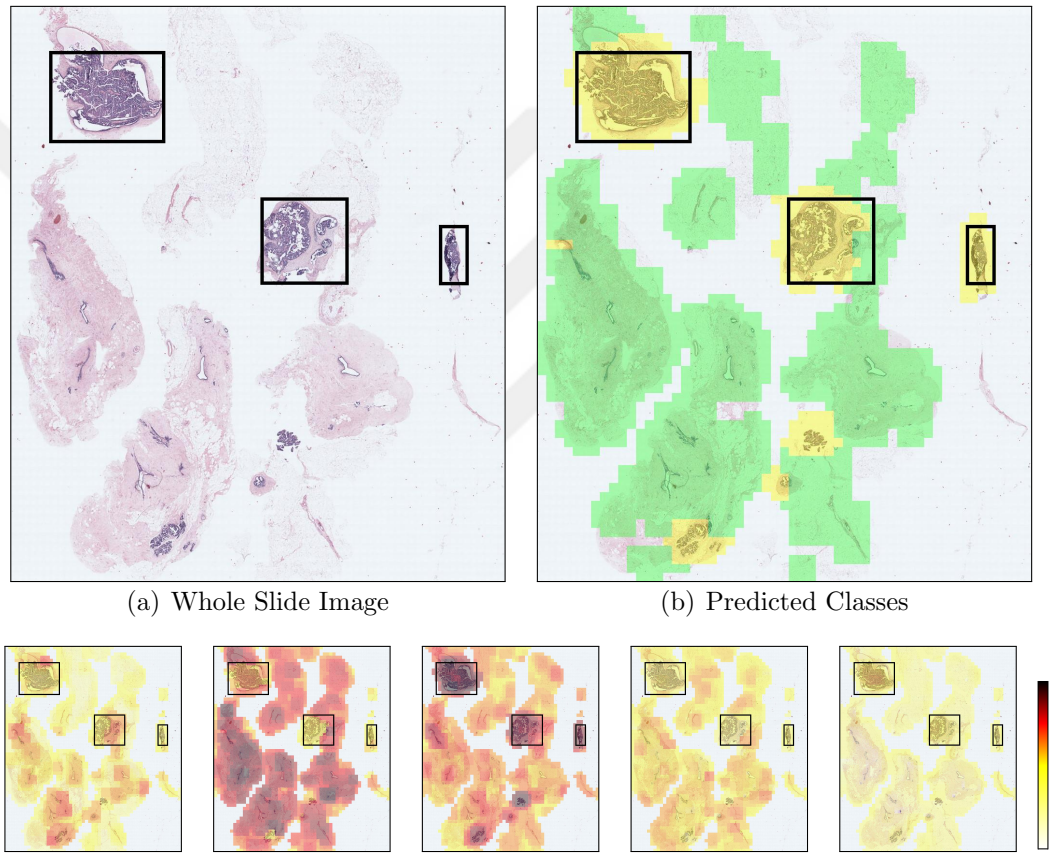
		Predicted				
		NP	P	ADH	DCIS	INV
True	NP	0	5	3	3	0
	P	0	15	30	13	4
	ADH	3	20	32	10	1
	DCIS	0	5	22	42	6
	INV	0	0	2	10	10

Table 4.8: Class-specific statistics on the performance of ROI-level classification. The number of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) are given. Precision, recall (also known as true positive rate and sensitivity), false positive rate (FPR), and specificity (also known as true negative rate) are also shown.

Class	TP	FP	FN	TN	Precision	Recall/ Sensitivity	FPR	Specificity
NP	0	3	11	222	0.00	0.00	0.01	0.99
P	15	30	47	144	0.33	0.24	0.17	0.83
ADH	32	57	34	113	0.36	0.48	0.34	0.66
DCIS	42	36	33	125	0.54	0.56	0.22	0.78
INV	10	11	12	203	0.48	0.45	0.05	0.95

only the 1200×1200 subwindows within that region were used for quantitative evaluation. We used the following protocol for predicting a label for this ROI by using its subwindows. First, we assigned the class that had the highest score as the diagnostic label of each subwindow. Then, we used a classification threshold on these scores to eliminate the ones that had low certainty. Finally, we picked the most severe diagnostic label among the remaining subwindows as the label of the corresponding ROI. If a slide-level grading is desired, the connected components formed by the subwindows that pass the classification threshold can be found, and the most severe diagnosis can be used as the diagnostic label of that slide. The components also provide clinically valuable information as one may want to localize all diagnostically relevant regions that may belong to different classes.

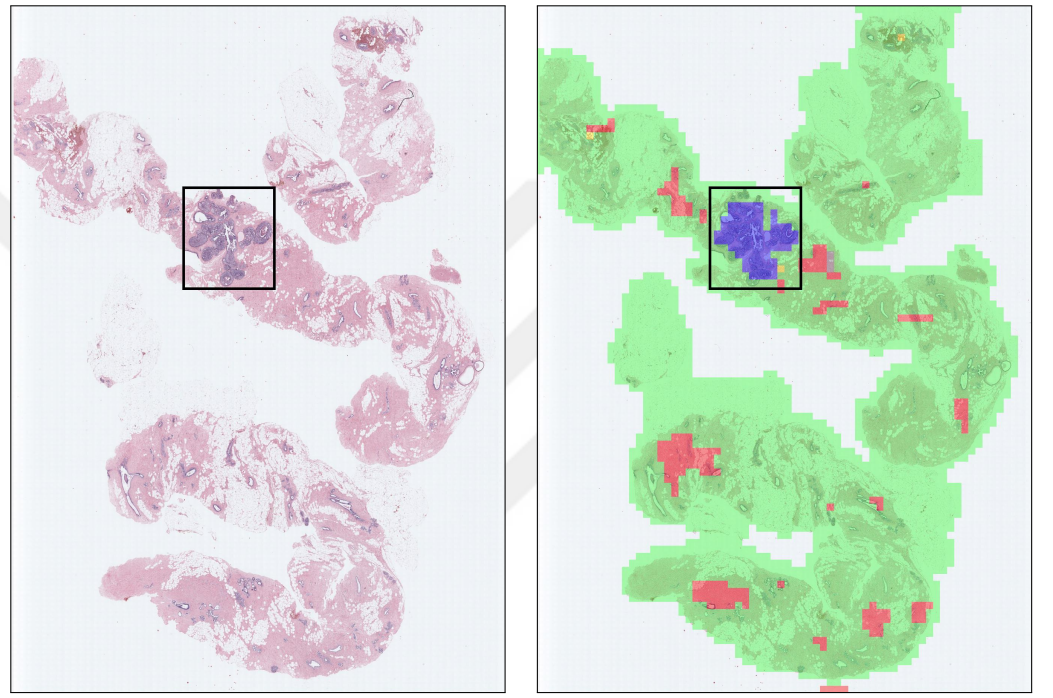
We evaluated different parameter settings for the protocol described above. The best results were obtained when the classification threshold was 0.7. Tables 4.7 and 4.8 summarize the classification results. Among the five classes, namely non-proliferative changes only (NP), proliferative changes without atypia (P), atypical ductal hyperplasia (ADH), ductal carcinoma in situ (DCIS), and invasive



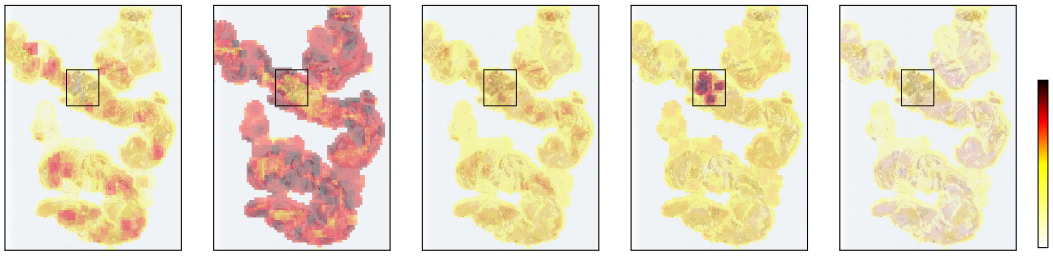
(a) Whole Slide Image (b) Predicted Classes

(c) NP (d) P (e) ADH (f) DCIS (g) INV

Figure 4.3: Whole slide ROI-level classification example for a case with ADH. First row, from left to right: original image; each 1200×1200 window is colored according to the class with the highest score (NP as red, P as green, ADH as yellow, DCIS as purple and INV as gray), second row from left to right: scores for individual classes using the color map show on the right. The consensus ROIs are shown using black rectangles. The consensus diagnosis for this case is atypical ductal hyperplasia.



(a) Whole Slide Image (b) Predicted Classes



(c) NP (d) P (e) ADH (f) DCIS (g) INV

Figure 4.4: Whole slide ROI-level classification example for a case with DCIS. First row, from left to right: original image; each 1200×1200 window is colored according to the class with the highest score (NP as red, P as green, ADH as yellow, DCIS as purple and INV as gray), second row from left to right: scores for individual classes using the color map show on the right. The consensus ROI is shown using a black rectangle. The consensus diagnosis for this case is ductal carcinoma in situ.

cancer (INV), we observed that the classifier could predict P, ADH, DCIS, and INV better than NP. In spite of the upsampling, most of the NP cases were incorrectly labeled as P, followed by ADH and DCIS. Precision values indicated better performance for DCIS and INV, followed by ADH and P. Recall values for P indicated a large number of missed cases; most were misclassified as ADH and a comparatively smaller number were misclassified as DCIS. ADH and DCIS were more successfully captured, with DCIS having a relatively smaller false positive rate compared to ADH where the classifier incorrectly assigned a class label of ADH to a large number of cases associated with P and a smaller number of cases associated with DCIS. The classifier could detect 10 out of the 22 INV cases correctly and 10 of the misclassified 12 cases were labeled as DCIS, which was not an unexpected result given that most cases labeled as INV also included DCIS in their pathology reports.

Even though slide-level predictions achieved precision values up to 81%, ROI-level quantitative accuracy appeared to be lower than human performance. The main cause of the ROI-level predictions counted as errors was the difficulty of the multi-class classification problem by using weakly-labeled learning from pathologists' viewing records. For example, the multi-label training sets with INV, DCIS, or ADH as the most severe diagnosis often also included other classes, and the candidate ROIs that were included in the bags that corresponded to these multi-label sets covered diagnostically relevant regions that belonged to the full continuum (P, ADH, low-grade DCIS, high-grade DCIS, etc.) of histologic categories. Unfortunately, there is no comparable benchmark that studied these classes in the histopathological image analysis literature where discrimination of classes such as ADH and DCIS was intentionally ignored as being too difficult even in fully supervised settings and when manually annotated ROIs were used for training [8, 53]. These classes are also often the most difficult to differentiate even by experienced pathologists using structural cues, and this was particularly apparent for our data set, as well [4, 37]. The proposed classification setting was powerful enough to work with generic off-the-shelf features that were not specifically designed for breast pathology. Figure 4.3 and Figure 4.4 present ROI-level classification examples. In general, the multi-class classification within the whole

slide and the localization of regions with different diagnostic relevance appeared to be more accurate compared to the numbers given in quantitative evaluation.

4.5 Discussion

This chapter presented a study on multi-class classification of whole slide breast histopathology images. Contrary to the traditional fully supervised setup, where manually chosen image areas and their unambiguous class labels are used for learning, we considered a more realistic scenario involving weakly labeled whole slide images where only the slide-level labels were provided by the pathologists. The uncertainty regarding the correspondences between the particular local details and the selected diagnoses at the slide level was modeled in a multi-instance multi-label learning framework, where the whole slide was treated as a bag, the candidate ROIs extracted from the screen coordinates as part of the viewing records of pathologists were used as the instances in this bag, and one or more diagnostic classes associated with the slide in the pathology form were used as the multi-label set.

Training and test data obtained through various combinations of three pathologists' recordings were used to evaluate the performances of four different multi-instance multi-label learning algorithms on classification of diagnostically relevant regions as well as whole slide images as belonging to 5 or 14 diagnostic categories. Quantitative evaluation of 5-class slide-level predictions resulted in average precision values up to 78% when individual pathologist's viewing records were used and 81% when the candidate ROIs and the class labels from all pathologists were combined for each slide. Additional experiments showed slightly lower performance for the more difficult 14-class setting. We also illustrated the use of classifiers trained using slide-level information for multi-class prediction of ROIs with different diagnostic relevance.

We would like to note that the 240 slides in our data set were selected to include the full range of cases, and with more cases of ADH and DCIS than in

Table 4.9: Kappa scores computed in different scenarios on the 5-class classification problem involving all pathologists who partially or fully evaluated the whole slide images in the data set. Scenarios include participant pairwise agreement, participant consensus agreement, and each data subset with a particular reference diagnosis for 5 classes.

	Min	Max	Mean	Std
Participant vs. participant	0.161	0.822	0.476	0.100
Participant vs. consensus	0.357	0.777	0.551	0.097
Non-proliferative changes only	-0.023	1.000	0.476	0.213
Proliferative changes	0.139	0.792	0.454	0.158
Atypical ductal hyperplasia	0.000	0.733	0.380	0.155
Ductal carcinoma in situ	0.340	0.961	0.662	0.141
Invasive carcinoma	0.535	1.000	0.908	0.097

typical clinical practice, this image cohort is diagnostically more difficult. We computed Kappa scores [54] involving all of the pathologists who participated in this research to emphasize the difficulty of the multi-class classification problem using whole slide images compared to commonly used scenarios where the focus is on carefully selected ROIs. All pairwise agreements and each pathologist's individual agreement with the consensus diagnosis in both 5-class and 14-class settings are shown in Table 4.9 and 4.10, respectively. The tables also include the Kappa values for each data subset with a particular reference diagnosis. Additionally, the classifiers used were trained only using weakly labeled data at the slide level, where the number of training samples could be considered very small for such a multi-class setting. Given the difficulty and the novelty of the learning and classification problems in this study, the results provide very valuable benchmarks for future studies on challenging multi-class whole slide classification tasks where collection of fully-supervised data sets is not possible.

Table 4.10: Kappa scores computed in different scenarios on the 14-class classification problem involving all pathologists who partially or fully evaluated the whole slide images in the data set. Scenarios include participant pairwise agreement, participant consensus agreement, and each data subset with a particular reference diagnosis for 14 classes.

	Min	Max	Mean	Std
Participant vs. Participant	0.163	0.625	0.408	0.081
Participant vs. Consensus	0.294	0.670	0.477	0.081
Non-proliferative changes only	-0.029	1.000	0.434	0.280
Fibroadenoma	-0.017	1.000	0.624	0.421
Intraductal papilloma w/o atypia	0.000	0.000	0.000	0.000
Usual ductal hyperplasia	-0.100	0.792	0.159	0.216
Columnar cell hyperplasia	-0.108	0.545	0.116	0.181
Sclerosing adenosis	-0.050	1.000	0.375	0.347
Radial scar complex sclerosing lesion	-0.026	1.000	0.151	0.315
Flat epithelial atypia	-0.087	1.000	0.212	0.256
Atypical ductal hyperplasia	-0.031	0.773	0.343	0.171
Intraductal papilloma with atypia	-0.071	1.000	0.452	0.307
Atypical lobular hyperplasia	-0.071	1.000	0.320	0.307
Ductal carcinoma in situ	0.308	0.921	0.653	0.144
Lobular carcinoma in situ	0.000	1.000	0.437	0.356
Invasive carcinoma	0.535	1.000	0.908	0.097

Chapter 5

Deep Feature Representations for Variable-sized ROIs in Whole Slide Images

The chapter introduces a new method for generating deep feature representations for variable-sized diagnostically relevant regions in a whole slide image either from patch-level or pixel-level information extracted from a fine-tuned deep network, weighted by each class specific prediction obtained from the same network. The diagnostically relevant regions in a slide generally differ in size greatly, and the number of regions in one whole slide image can be substantially different from another whole slide image. Hence, training of classifiers can be performed more effectively when the tissue structures used in learning are in isolation and have no ambiguity in their diagnostic labels. This setting requires the pathologists to annotate the regions of interest (ROI) corresponding to the diagnostically relevant regions in whole slide images and associate each ROI with one of the diagnostic labels. Performance improvements on the multi-class classification of malignant regions can be achieved when isolated ROIs with the associated labels are used in the training of the state-of-the-art deep learning based models.

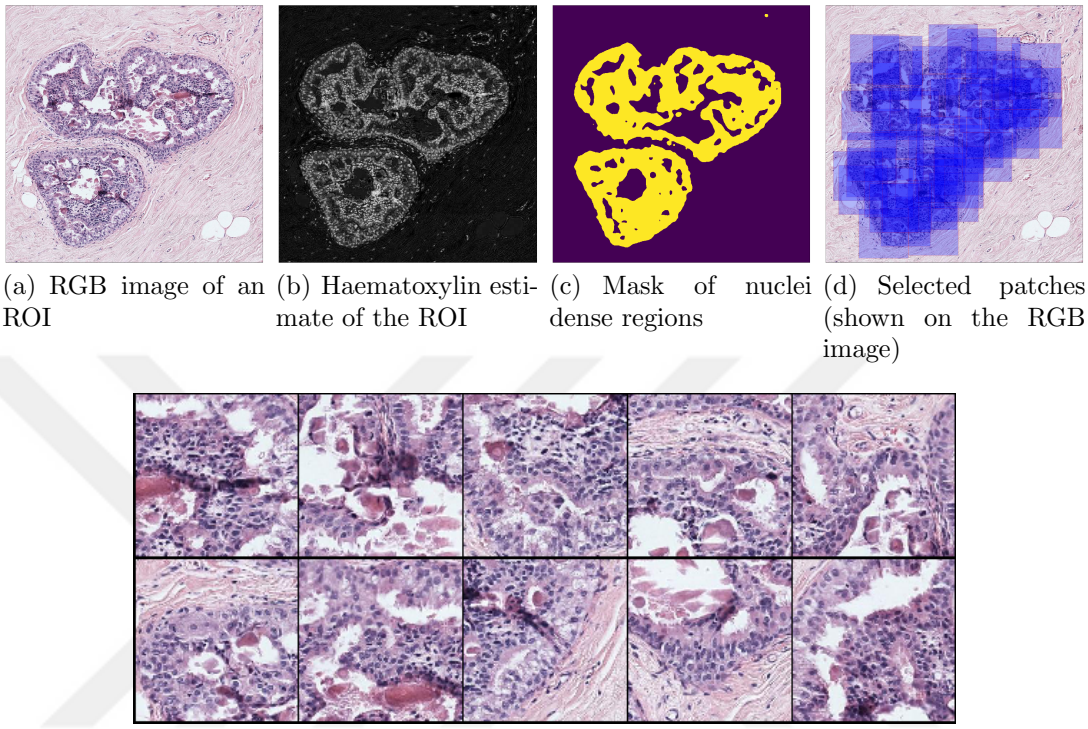
Convolutional neural networks (CNNs) have had great success in several different domains including histopathological image analysis [18–20]. CNNs require the input image to be of specific size, typically very small, due to the computational limitations. This poses a problem for the classification task of the ROIs in histopathology images, due to them having arbitrarily different sizes in high-resolution. CNNs trained on cropped patches from an ROI may not be able to preserve the contextual information within the ROI, and using a resized version of the ROI as input to CNNs results in structural information loss. Earlier works in the histopathological image analysis using deep networks focused on binary (benign vs. malignant tissue) classification problem [21–23, 26]. Classifying a tissue as one of the many subcategories of breast cancer holds a more clinical significance compared to as benign vs. malignant. There have been also some works that performed multi-class classification of breast histopathology images using CNNs [28–30]. Additionally, the representational capabilities of deep networks were also investigated in several works in the field of histopathology image analysis [31–35]. However, the majority of the earlier works perform only patch-level classification. Some use simple fusion-based approaches such as majority voting and aggregation of patch-level probabilities to predict the slide-level diagnosis while others learn slide-level feature representations from only small fixed-sized patches. We propose a feature generator network that is used to learn representations for ROIs of any shape or size in whole slide images by aggregating the information extracted from a deep convolutional network fine-tuned at patch-level.

The main contribution of this study is the introduction of a deep feature generation method for ROIs in whole slide images regardless of their shape or size that preserves the local as well as the contextual information within the ROI. We fine-tune a deep convolutional network on the patches extracted from ROIs using the ROI-level diagnostic labels. A feature vector of a patch from an ROI of an unseen whole slide image is then extracted from specific layer activations of the fine-tuned deep network. We preserve the class specific information of a patch by applying each class output from the softmax layer of the same network

as direct coefficients on the feature vector of the patch. The feature representations of the patches are then aggregated to obtain the feature representation of the ROI from which the patches were previously extracted. We demonstrate two separate deep feature vector extraction methods applicable to any type of convolutional deep network architecture to evaluate the effectiveness of the proposed ROI-level feature generation model from both pixel-level and patch-level information. The final ROI-level feature representation simultaneously preserves the local information contained in the patches and encodes the class distributions of the patches within the ROI. We show that the proposed deep feature generation method can be easily applied to any variable-sized ROIs and we provide multi-class classification settings with ROI-level classifiers trained on the feature representations to demonstrate the superiority of our method compared to the existing ones at ROI-level classification of breast histopathology images. In addition, we present an extension to the proposed ROI-level model to demonstrate its slide-level classification performance.

5.1 Patch-level Deep Network Training

The variable structure of the ROIs poses a great challenge for the state-of-the-art deep convolutional architectures that require fixed-sized small patches that preferably contain structural and contextual information sufficient for classification. Therefore, whole slide image analysis, even in the presence of annotated ROIs within slides, faces the limitation of convolutional networks due to ROIs having variable shapes and sizes. Therefore, we need techniques to extract potentially informative patches to feed to CNNs which in return, can be used to generate deep feature representations for ROIs.



(a) RGB image of an ROI

(b) Haematoxylin estimate of the ROI

(c) Mask of nuclei dense regions

(d) Selected patches (shown on the RGB image)

(e) Example extracted patches (shown separately)

Figure 5.1: Patch selection process shown on an example ROI.

5.1.1 Identification of Patches from ROI

We represent an ROI with a set of variable number of fixed-sized patches, directly extracted from the ROI. The extracted patches need to be informative and diverse enough to represent the structural and contextual information in the ROI. We use the haematoxylin channel estimated from the RGB image using a color deconvolution procedure [42] to locate the nuclei dense regions. A non-parametric Parzen density estimate was built in the image domain by applying a Gaussian window on the haematoxylin values of the pixels, and a threshold was applied to this estimate to eliminate the regions with little to no nuclei. The points inside the resulting nuclei region correspond to the center points of candidate patches. We extract potentially informative patches from the ROI using these points and achieve diversity between patches by placing a constraint which imposes that two patches should never overlap within a margin. The number of patches to extract within an ROI is automatically determined by the size of its nuclei region. The patch selection process is presented in Figure 5.1 on an example ROI.

5.1.2 CNN Training on Patches

Due to the scarcity and the limited availability of the histopathological images, we opt to use a pre-trained network and fine-tune its classification parameters in the fully-connected layers on the patches extracted from the ROIs using the associated diagnostic labels. Please note that the parameters in the convolutional layers are kept frozen during the fine-tuning of the network. Despite our efforts to train the parameters of the convolutional layers by training the network from scratch, the performance of the network suffered a lot due to the very limited number of available data for training.

5.2 ROI-level Deep Feature Representations

We propose a deep feature generator network to learn deep feature representations for variable number of variable sized ROIs in whole slide images. This generator network defines the deep feature representation of an ROI as the aggregation of the feature representations of its patches through average pooling. A patch-level feature representation is defined as the aggregation of its deep feature vectors, extracted from the activations of the fine-tuned deep network and weighted by each class specific output from the same network. This approach is related to the super-vector coding method [55] where the codebook corresponds to the feature vector of the patch extracted from the deep network and the coefficients correspond to the patch-level class probabilities from the same network. In contrast to the vanilla super-vector coding method where the coefficients are in one-hot representation, the coefficients in our model correspond to the class probabilities from the deep network which can have values between 0 and 1, always summing to 1.

We propose two different methods for extracting the deep feature vectors, i.e. the codebooks, from the fine-tuned network. The first method uses the patch-level CNN penultimate layer activations as the feature vector of a given patch and the second method incorporates pixel-level information by operating over the

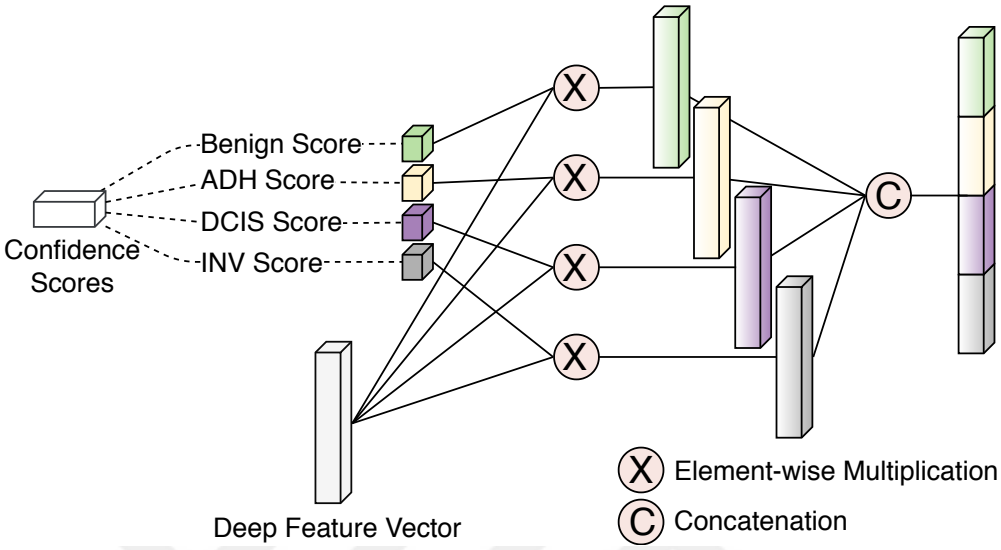


Figure 5.2: The set of operations are demonstrated on the feature vector extracted from a fine-tuned deep network and on the prediction scores of the same network for an input patch to obtain its feature representation.

convolutional layer activations to form the feature vector of the patch. The rest of the operations that leads to the ROI-level feature representation follows the same procedure for both methods. The feature vector of the patch is weighted by each class probability and the resulting vectors are concatenated to form the patch-level feature representation. The ROI-level feature representation is defined as the averaged feature representations of the patches extracted from the ROI. A visualization of the operations leading to the feature representation of a patch using the deep feature vector of the patch from the network activations and the confidence scores from the network outputs is presented in Figure 5.2.

5.2.1 ROI-level Feature Representation from Weighted Patch-level Penultimate Layer CNN Features

The first method extracts the feature vector of a patch from the penultimate layer of the deep network. A patch from the automatically identified set of patches of an ROI is fed to the fine-tuned deep network to extract its feature vector from the penultimate layer. This process is repeated for each patch to obtain the final patch-level feature representation, which is the concatenated

penultimate layer features weighted by each class probability from the final softmax layer of the same network. Finally, the feature representation of the associated ROI is computed by averaging over the feature representations of its patches. More specifically, let R denote the feature representation of an ROI and $\{r_1, r_2, \ldots, r_M\} \in R$ denote the patches extracted from the ROI, with M denoting the number of patches. The CNN feature vectors of the extracted ROI patches are $\{\hat{\phi}(r_1), \hat{\phi}(r_2), \ldots, \hat{\phi}(r_M)\}$, and the class-specific softmax probabilities of a patch, r_m, are given as $\{s_m^1, s_m^2, \ldots, s_m^K\}$ where K is the number of class labels and $\sum_{k=1}^{K} s_m^k = 1$. We compute the feature representation of a patch, $\phi(r_m)$, as

$$\phi(r_m) = \left[s_m^1 \hat{\phi}(r_m), s_m^2 \hat{\phi}(r_m), \ldots, s_m^K \hat{\phi}(r_m)\right] \tag{5.1}$$

Consequently, we compute the corresponding ROI feature representation, $\phi(R)$, as

$$\phi(R) = \left[\frac{1}{M}\sum_{m=1}^{M} s_m^1 \hat{\phi}(r_m), \frac{1}{M}\sum_{m=1}^{M} s_m^2 \hat{\phi}(r_m), \ldots, \frac{1}{M}\sum_{m=1}^{M} s_m^K \hat{\phi}(r_m)\right]. \tag{5.2}$$

5.2.2 ROI-level Feature Representation from Weighted Pixel-level Hypercolumn CNN Features

The second method investigates the activations in several convolutional layers to extract the feature vector from the hypercolumns of the network. The earlier activations provide low-level information such as color, texture and shape features while the activations in the final convolutional layers encode contextual information related to the input image [56]. The hypercolumn CNN features combine the low-level as well as the high-level features to represent a pixel in the image. The hypercolumn feature representation of a pixel is obtained by concatenating the CNN activations at the corresponding pixel location through the layers in the convolutional network [57]. However, the number of activations at each channel of a convolutional layer decreases towards the end of the convolutional network due to the max pooling operations in the network. Simple solutions such as upsampling [57] and skip-connections [58] through the convolutional layers were employed to cope with this problem which were useful for the segmentation problems where pixel-level information is particularly important.

Our aim is to extract a patch-level feature representation from the CNN hypercolumn features. The naive concatenation of the hypercolumn features would output a huge feature vector which would be prone to overfitting and would lead to computational problems. Even on a mildly deep network such as the VGG16 network [59], the concatenation of the CNN hypercolumn features results in a feature vector with size over two hundred billion. We follow a procedure which involves operations on the select layers of the convolutional network to form a feature representation to an input patch using the pixel-level CNN hypercolumn features. We select a subset of the convolutional layers from the set of layers placed before a max pooling layer since these layers encode increasing representational capacity as the number of activations do not change but the convolutional operations increase as the input is passed through each layer consecutively. The operations to obtaining patch-level convolutional feature representation from the pixel-level CNN hypercolumn feature representations are as follows. First, we take only the maximum activations of each convolutional channel of a layer by turning off the remaining activations. Then, we combine the top activations by summing over the convolutional channels to obtain the aggregate activation map of the associated convolutional layer. The resulting aggregate activation map preserves the most prominent responses that contain information from different local structures activated by various convolutional kernels. The maps in different layers have varying ranges, i.e. maps from the earlier layers have fewer number of convolutional channels which results in smaller activation values in the associated maps. Therefore, we threshold the aggregate activation map by the median to not only keep the most descriptive responses, but also normalize the activation maps across the convolutional layers. Finally, the resulting activation mask is vectorized and all of the vectorized masks from the select convolutional layers are concatenated to obtain the pixel-level CNN hypercolumn feature vector of the input patch. The activations in the convolutional channels at one of the select layers and the corresponding top activation maps as well as the aggregate activations mask with the resulting feature vector following these operations on an example patch extracted from an ROI is presented in Figure 5.3.

More specifically, given a set of convolutional activations $A_c, c \in \{1, \ldots, C\}$ at

convolutional layer l, $l \in \{1, \ldots, L\}$ of an input patch r_m, where C denotes the number of convolutional channels and L denotes the number of select layers in the convolutional network. We aggregate the convolutional activations as

$$\hat{A}_c(x, y) = \begin{cases} A_c(x^*, y^*) & if\ (x^*, y^*) = \arg\max_{(x,y)} A_c \\ 0 & otherwise \end{cases} \tag{5.3}$$

$$A_{AGG} = \sum_c \hat{A}_c. \tag{5.4}$$

Then, the activation mask is computed as

$$A_{MASK} = \{(x, y) | A_{AGG}(x, y) \geq median(A_{AGG})\}. \tag{5.5}$$

Finally, the patch-level CNN hypercolumn feature vector of the patch, r_m, is obtained as

$$\hat{\phi}_H(r_m) = \left[vec(A^{(1)}_{MASK}), vec(A^{(2)}_{MASK}), \ldots, vec(A^{(L)}_{MASK})\right] \tag{5.6}$$

where $vec(\cdot)$ denotes vectorization operation of a given matrix. We repeat the above procedure to compute the pixel-level CNN hypercolumn feature vector for every patch extracted from an ROI. The final feature representation of the ROI is obtained following the same procedure described in Section 5.2.1, but now with the pixel-level hypercolumn feature vectors instead of the penultimate feature vectors.

As a pre-processing step prior to the classification, we apply Principle Component Analysis (PCA) [60] to remove redundancy and noise from the feature representations of the ROIs.

5.3 Classification

We perform both ROI-level and slide-level classification from the deep feature representations of ROIs. The slide-level classification involves additional methods to extend the ROI-level classification results to slide-level.

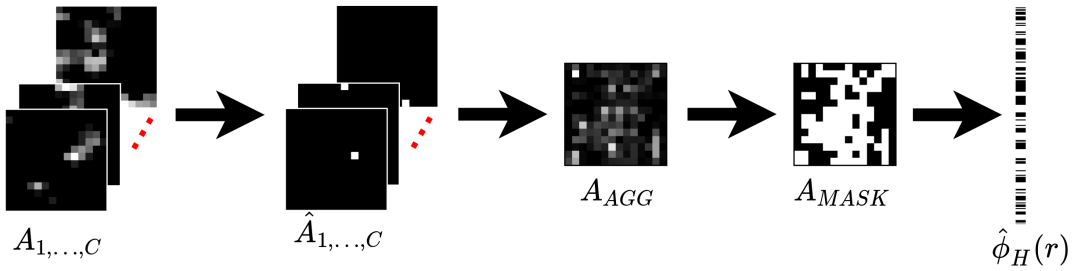


Figure 5.3: The process of obtaining a feature vector from the convolutional channels at a select layer of a fine-tuned deep network for an input patch.

5.3.1 ROI-level Classification

The ROI-level feature representations from the training set are used to train a multi-layer perceptron (MLP) to perform multi-class classification on unseen ROIs in the test sets whose feature representations are also extracted with the same procedures.

5.3.2 Slide-level Classification

Whole slide images may contain several diagnostically relevant regions. A whole slide is associated with the most severe diagnosis present in one of these regions. Therefore, slide-level classification is a challenging task that involves detecting the diagnostically relevant regions inside the slide and associating them with one of the diagnostic labels. In this study, we use a sliding window approach to infer the label of the whole slide from the predictions of the regions inside the associated windows. Each window overlaps with one another both horizontally and vertically. We can consider a window as a candidate ROI which may or may not correspond to a diagnostically relevant region. A set of preprocessing steps are applied on these candidate ROIs to prune the redundant ones to speed up the classification performance. A candidate ROI is removed if it overlaps with the background region inside the whole slide with more than a threshold or it contains little to no nuclei region. Then, the remaining ROIs are merged into a larger ROI if they overlap with each other. The pruning step discards

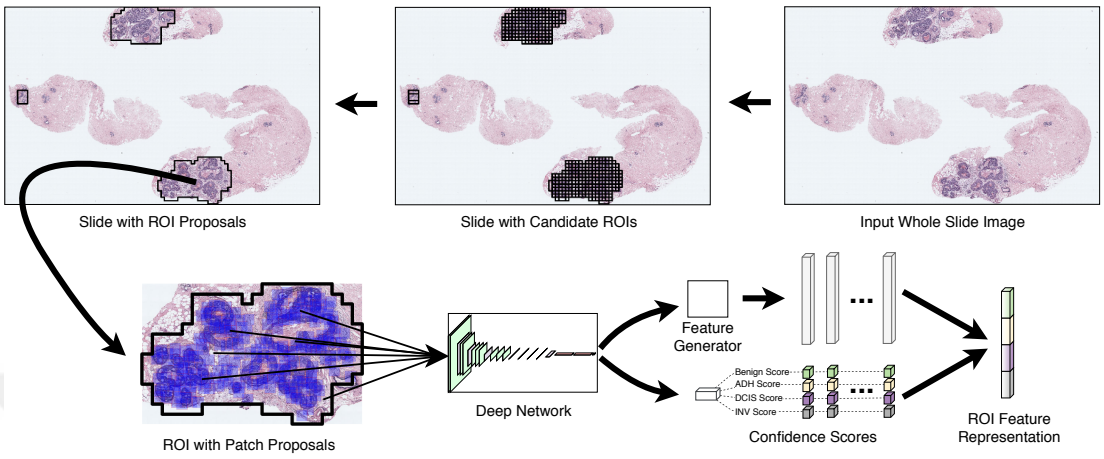


Figure 5.4: The ROI-level feature representation steps. The ROI feature representation of an ROI proposal obtained from an input whole slide image is computed from the properties of a fine-tuned deep convolutional network.

tissue that are not diagnostically relevant and the merging step helps preserving the contextual information of the tissue. We refer the remaining ROIs after the preprocessing steps as ROI proposals. The proposed feature representations of an ROI proposal is then computed from its associated patch-level representations using the properties of the fine-tuned CNN as described in Section 5.2.1 and 5.2.2. The process of obtaining the feature representation of an ROI proposal obtained from a whole slide image is presented in Figure 5.4. Subsequently, we use the same classifiers in Section 5.3.1 to compute the confidence scores and to predict the diagnostic labels of the ROI proposals. We repeat these steps for every ROI proposal in the slide. Finally, the label of the slide is determined by employing two different decision aggregation methods: the Max-pooling method [6] by pooling the distribution of the class probabilities of the ROI proposals and the decision fusion method [61] by exploiting the class frequency histograms of the ROI proposals in the slide.

5.4 Experimental Results

The data set of 437 ROIs belonging to one of the 4 classes (benign without atypia (Benign), atypical ductal hyperplasia (ADH), ductal carcinoma in situ (DCIS),

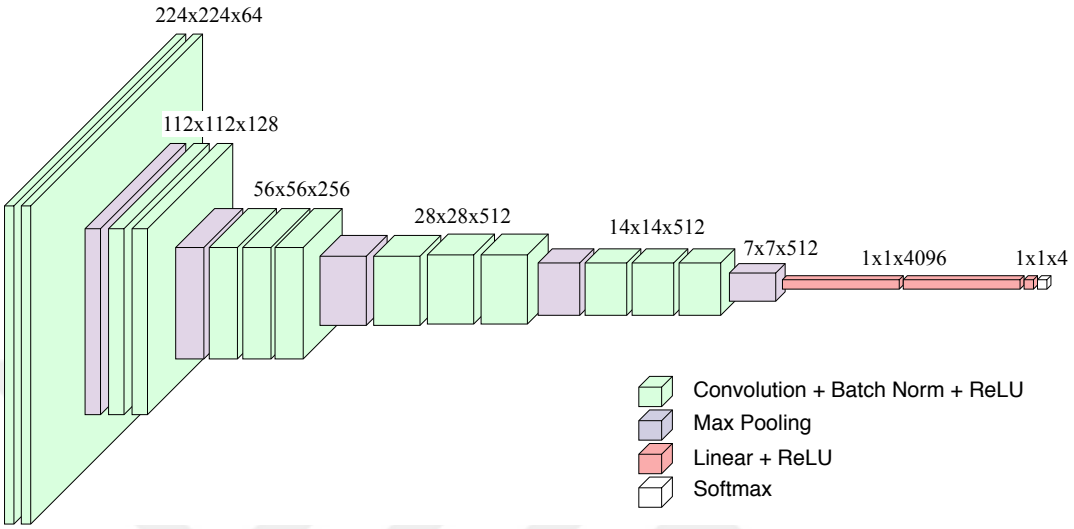


Figure 5.5: The architecture of the VGG16 network with batch normalization.

and invasive cancer (INV)) is split into two equal sized sets containing slides from different patients for training and test. The slide-level class distribution between the two sets are kept as close as possible. The distributions of the slides and the associated ROIs in the training and the test sets are presented in Table 5.1. Please note that the data set split is based on the slide-level class distribution and it does not take the ROI-level class distribution into account. Also note that the ROIs in the data set have considerable amount of variability in size as the largest ROI in the set is over 2000 times larger than the smallest ROI. In our experiments, we used the VGG16 network [59] as the base CNN architecture due to its relatively large depth and representational capabilities. However, the feature extraction process is not specific to the network architecture and can be applied to any convolutional network. The specific CNN architecture, VGG16 network with batch normalization applied before ReLU activations, is presented in Figure 5.5.

To improve the generalization performance of the VGG16 network, we applied random rotation, random horizontal and vertical flipping, random perturbations on the Hue channel in the HSV image domain as part of the data augmentation routine. We fine-tuned the network on the augmented training set using cross-entropy loss, leaving a small portion of the training set for validation. We used batches of 12 patches, employed Adam optimizer with a learning rate set to $1e{-}4$,

Table 5.1: The class distribution of the slides and the ROIs in the training and the test sets.

		Benign	ADH	DCIS	INV	Total
Slide	Training Set	34	35	41	10	120
	Test Set	22	48	38	12	120
ROI	Training Set	60	58	85	17	220
	Test Set	37	81	80	19	217

and fixed the dropout to 0.75 on the fully connected layers of the VGG16 network. The VGG16 network in our experiments had 13 convolutional layers and three fully-connected layers where the last layer outputs predictions for each of the K classes through a softmax activation function. For the notation simplicity, we denote the convolutional layers as $conv - \{1_1, 1_2, 2_1, 2_2, 3_1, 3_2, 3_3, 4_1, 4_2, 4_3, 5_1, 5_2, 5_3\}$ and fully-connected layers as $fc - \{1, 2, 3\}$. The penultimate layer activations of a patch were directly extracted from $fc - \{2\}$ of the network while the hypercolumn CNN features of a patch were computed from the convolutional activations in $conv - \{3_3, 4_3, 5_3\}$.

5.4.1 ROI-level Classification Results

We used the fine-tuned VGG16 network to extract the feature representations of these ROIs. We employed two approaches to obtain the feature representations of the ROIs, called Penultimate-Feat-Weighted and Hypercolumn-Feat-Weighted, as described in Section 5.2.1 and Section 5.2.2, respectively. The ROI-level feature representations in the training data were used to train a 4-class multi-layer perceptron (MLP) classifier on a 3-fold cross validation setting. The feature representation of an unseen ROI follows the same feature extraction routine and the MLP classifier is used to predict the label of the ROI. We compared the performance of our proposed algorithms with the following methods

- Penultimate-Feat-Base: The patch-level features extracted from the penultimate layer of the CNN without weighting with the class probabilities, were averaged to obtain the feature representation of the associated ROI. An MLP classifier is trained from the feature representations of the ROIs

in the training set.

- Hypercolumn-Feat-Base: The patch-level CNN hypercolumn features of the extracted ROI patches from $conv - \{3_3, 4_3, 5_3\}$ without weighting with the class probabilities, were averaged to obtain the feature representation of the associated ROI. An MLP classifier is trained from the feature representations of the ROIs in the training set.

- Max-Pooling: The patch-level class probabilities from the final softmax layer of the network were pooled to obtain the class label of the associated ROI [6]. First, a classification threshold was applied on the probabilities to eliminate the predictions with low confidence. Then, the label of the overlapping patch parts was determined by taking the prediction with the largest probability of the overlapping patches. Finally, a frequency threshold was applied on the histogram of the remaining predictions to eliminate the labels with too little coverage. In our experiments, we used 0.50 as the classification threshold and set the frequency threshold to 0.25. The label of the ROI was selected as the most severe diagnostic label within the predictions of the patches.

- Decision-Fusion: The patch-level class probabilities from the final softmax layer of the network were summed up to create a class frequency histogram of the associated ROI [30]. The ROI class frequency histograms and the associated class labels were then used to train an MLP classifier.

- Y-Net: The segmentation of patches inside an ROI was used in conjunction with the patch-level probability mask on a modified version of U-Net [62], called Y-Net, to improve the segmentation performance of the network [20]. Please note that, this study included the same data set and the same data setting in their experiments, therefore we could directly compare the classification performance of Y-Net with our results.

We presented the comparison of the ROI-level classification performance of the methods in Table 5.2. Both of the proposed methods outperformed other methods in terms of the classification performance at ROI-level. The best accuracy of 0.691

Table 5.2: The comparison of ROI-level classification performance.

Method	Accuracy
Pathologists [4]	*0.700*
Max-Pooling [6]	0.548
Decision-Fusion [30]	0.649
Y-Net [20]	0.625
Penultimate-Feat-Base	0.622
Hypercolumn-Feat-Base	0.585
Penultimate-Feat-Weighted	0.668
Hypercolumn-Feat-Weighted	**0.691**

was obtained when using the weighted patch-level CNN hypercolumn feature vectors for the ROI feature representations, Hypercolumn-Feat-Weighted, and it is followed by the ROI representations obtained from the weighted patch-level CNN penultimate layer feature vectors, Penultimate-Feat-Weighted, with an accuracy of 0.668. When we compared Penultimate-Feat-Weighted and Hypercolumn-Feat-Weighted to the corresponding vanilla feature representations, Penultimate-Feat-Base and Hypercolumn-Feat-Base, respectively, we observed that the former had an edge over the latter in contrast to their weighted counterparts. The reason could be due to the fact that the penultimate layer and the final softmax layer were located close to each other and they encoded similar properties so that the improvement was limited. On the other hand, the hypercolumn features were obtained from the convolutional units which could encode different characteristics of the patches, and therefore, could boost the performance of the corresponding vanilla feature representation more dramatically when combined with the encoded properties from the softmax layer. Additionally, between the two pooling approaches, Max-Pooling and Decision-Fusion, the latter outperformed the former due to it learning a classifier from the pool of patch-level class probabilities instead of performing simple pooling operations from the probabilities using predefined thresholds.

The confusion matrix and the class specific statistics of the classification performance of the Penultimate-Feat-Weighted features are given in Table 5.3 and 5.4, respectively. Out of the four classes, Benign without Atypia (Benign), Atypical ductal hyperplasia (ADH), Ductal carcinoma in situ (DCIS), and Invasive cancer (INV), the classifier could predict ADH and DCIS better than Benign and

INV. The classifier was able to predict the ROIs with ADH 56 out of 81 times, misclassifying only 15 of them as Benign and 10 of them as DCIS. Additionally, the classifier achieved the best Precision value on ADH, only misclassifying 19 out of the 75 ROIs as ADH. ROIs with DCIS were best captured by the classifier, misclassifying only 13 out of the 80 ROIs. The Precision performance on DCIS was behind ADH, with the classifier favoring DCIS in more cases. Majority of the incorrectly labeled ROIs with Benign were predicted as ADH, followed by DCIS. The ROIs with INV were correctly predicted in seven ROIs compared to the 10 ROIs which were predicted as DCIS.

Table 5.3: Confusion matrix of Penultimate-Feat-Weighted for ROI-level classification.

		Predicted			
		Benign	ADH	DCIS	INV
True	Benign	15	12	8	2
	ADH	15	56	10	0
	DCIS	5	6	67	2
	INV	1	1	10	7

Table 5.4: Class-specific statistics on the performance of Penultimate-Feat-Weighted features for ROI-level classification. The number of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) are given. Precision, recall (also known as true positive rate and sensitivity), false positive rate (FPR), specificity (also known as true negative rate) and F-measure are also shown.

Class	TP	FP	FN	TN	Precision	Recall/ Sensitivity	FPR	Specificity	F-measure
Benign	15	21	22	159	0.417	0.405	0.117	0.883	0.411
ADH	56	19	25	117	0.747	0.691	0.140	0.860	0.718
DCIS	67	28	13	109	0.705	0.837	0.204	0.796	0.766
INV	7	4	12	194	0.636	0.368	0.020	0.980	0.467

The confusion matrix and the class specific statistics of the classification performance of the Hypercolumn-Feat-Weighted features are given in Table 5.5 and 5.6, respectively. The classifier was able to classify more INV cases correctly and Benign cases were also predicted with better accuracy with the cost of misclassifying a portion of the cases with ADH. Similarly, the classifier was able to more successfully predict the cases with INV, and a smaller number of cases with INV were misclassified as DCIS which would make sense given that a large number of cases with INV also involved DCIS in their pathology reports [63].

Table 5.5: Confusion matrix of Hypercolumn-Feat-Weighted for ROI-level classification.

		Predicted			
		Benign	ADH	DCIS	INV
True	Benign	24	5	7	1
	ADH	23	50	8	0
	DCIS	5	7	66	2
	INV	0	1	8	10

Table 5.6: Class-specific statistics on the performance of Hypercolumn-Feat-Weighted features for ROI-level classification. The number of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) are given. Precision, recall (also known as true positive rate and sensitivity), false positive rate (FPR), and specificity (also known as true negative rate) are also shown.

Class	TP	FP	FN	TN	Precision	Recall/ Sensitivity	FPR	Specificity	F-measure
Benign	24	28	13	152	0.462	0.649	0.156	0.844	0.539
ADH	50	13	31	123	0.794	0.617	0.096	0.904	0.694
DCIS	66	23	14	114	0.742	0.825	0.168	0.832	0.781
INV	10	3	9	195	0.769	0.526	0.015	0.985	0.625

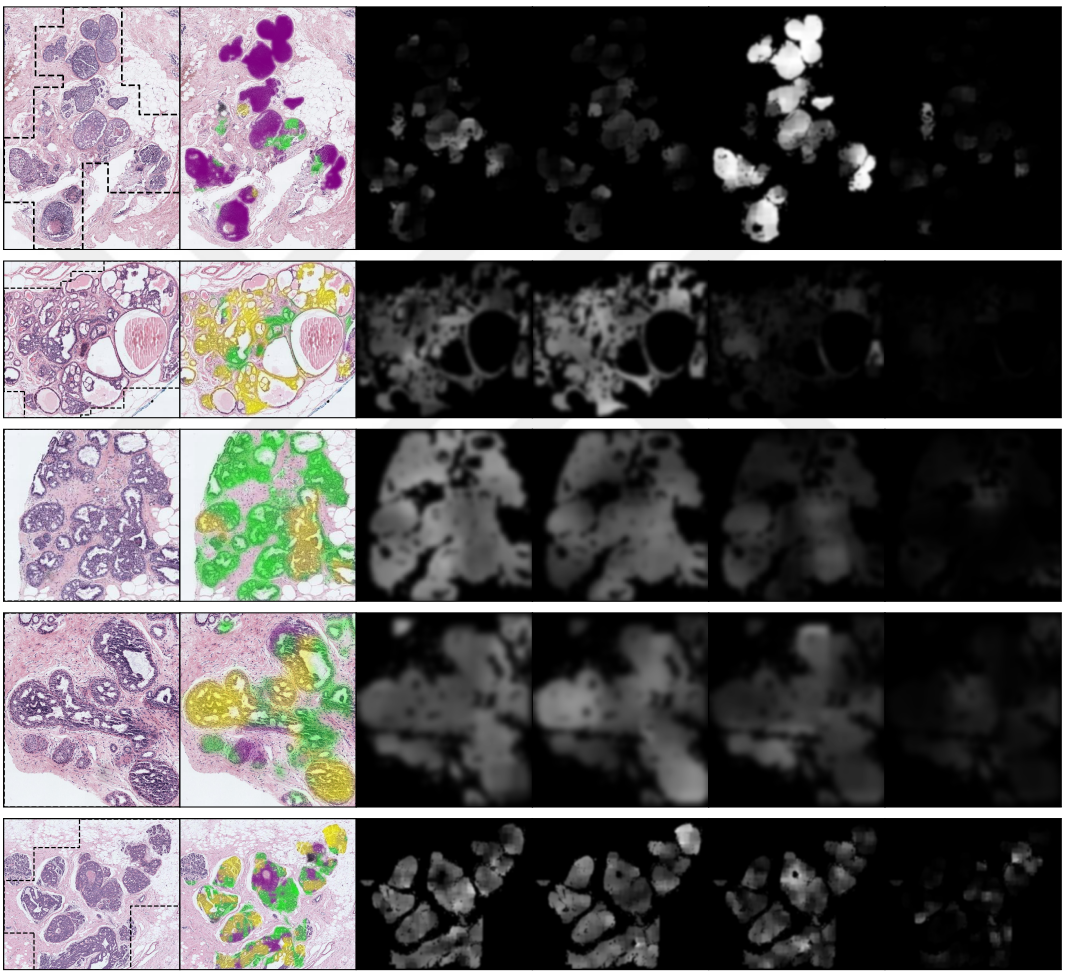
Classifiers trained on both of the proposed feature representations outperformed the other classifiers in the competition. In spite of our efforts to even the total number of patches extracted from each class during the fine-tuning of the network, the number of ROIs with Benign and especially with INV was very small compared to ADH and DCIS in our data set. Although the total number of the extracted patches were kept similar across the classes through upsampling from the minority classes, the diversity of the extracted patches varied greatly from one class to another due to the imbalanced number of ROIs across different classes which would explain the relatively poor performance on Benign and INV. Comparing the results from the classifiers trained on the two proposed feature representations, we can see the improvement of the inclusion of pixel-level information to the patch-level feature representation. The classifier learned the characteristics of the minority classes, Benign and INV, better with a small cost of misclassifying more ADH cases as Benign.

As described previously, various approaches were performed to predict the ROI-level label from the patch-level predictions. We present the patch-level CNN

predictions and class-specific scores from the fine-tuned VGG16 network on example ROI images in Figure 5.6. The patch-level CNN predictions for the ROI in the first row mostly involved DCIS as almost all patches within the ROI showed the strongest responses to DCIS as seen from the individual class scores. The methods involving the proposed feature representations, Penultimate-Feat-Weighted and Hypercolumn-Feat-Weighted, and the methods we used for comparison, Max-Pooling and Decision-Fusion, were able to correctly classify the ROI as DCIS. Similarly, most patches were diagnosed as ADH and some were also classified as Benign within the ROI in the second row. The proposed methods as well as the comparison methods correctly classified the ROI as ADH. The experimental results for the ROI on the third row were also similar in the sense that the proposed as well as the comparison methods could classify the ROI as Benign since patch-level predictions corresponded to the ROI-level consensus label, Benign. However, when the patch-level predictions from the CNN did not fully represent the consensus diagnosis of the ROI, the comparison methods performed poorly as seen on the ROIs in the fourth and the fifth rows. Both of the proposed methods were able to classify the ROI in the fourth row as ADH whereas only Hypercolumn-Feat-Weighted could correctly predict the class label of the fifth ROI. The consensus diagnoses were ADH and DCIS for the ROIs in the fourth and the fifth row, respectively. The visual and the quantitative experimental results showed that the predictions from the patch-level CNN outputs may not be representative enough to make ROI-level predictions. The proposed approaches were able to more successfully associate the correct diagnostic label with an ROI even when the patch-level diagnoses from the ROI did not fully represent the consensus label of the ROI.

5.4.2 Slide-level Classification Results

The slide-level classification of whole slide images follows the procedure from Section 5.3.2. The slides were processed within sliding windows of 600×600 pixels with an overlap of 200 pixels both horizontally and vertically. A window was removed from the set of candidate ROIs if more than half of its area fell in



(a) ROI (b) Predictions (c) Benign (d) ADH (e) DCIS (f) INV

Figure 5.6: Patch-level classification outputs from CNN softmax predictions with max-pooling on example ROIs. From left to right: RGB image of an ROI with consensus ROI shown in black dashed line; predicted classes from the network, Benign as green, ADH as yellow, DCIS as purple, INV as gray; scores for individual classes, Benign, ADH, DCIS and INV.

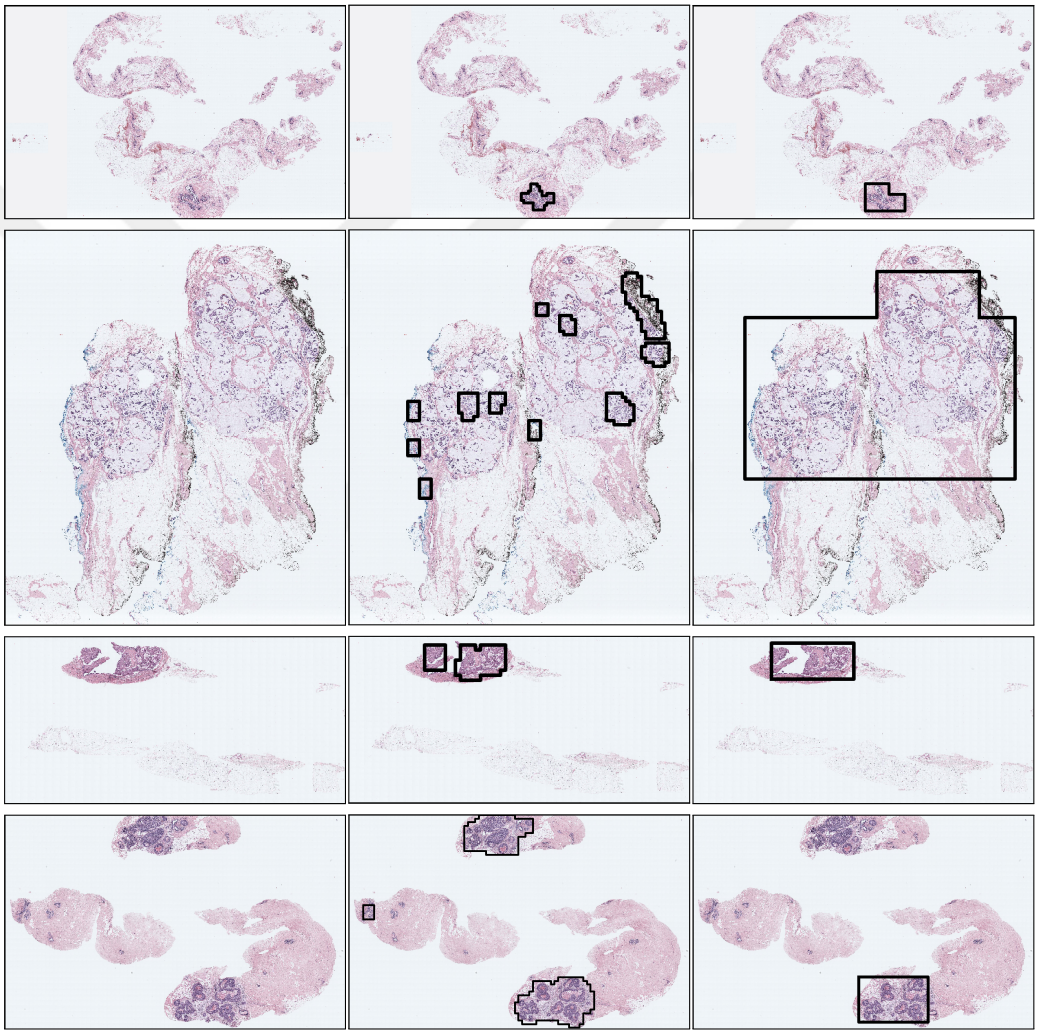
Table 5.7: The slide-level classification performance comparison.

ROI Predictions	Slide Pooling	Accuracy
Penultimate-Feat-Weighted	Max-Pooling	0.525
Penultimate-Feat-Weighted	Decision-Fusion	0.500
Hypercolumn-Feat-Weighted	Max-Pooling	0.533
Hypercolumn-Feat-Weighted	Decision-Fusion	0.533

the background region of the slide. Subsequently, the haematoxylin channel of the tissue patch inside a sliding window was used to determine if the window contained regions with nuclei, so that the windows that contained little to no nuclei regions were eliminated. This process followed the procedure that was described in detail in Section 5.1.1. Finally, regions within the remaining windows were merged into a larger region if the associated windows overlapped with each other, which we defined previously as ROI proposals. The properties of the sliding windows and the thresholds to window elimination and merging steps were set by maximizing the area of the intersection of the output ROI proposals and the consensus ROIs of the associated slide for each slide in the training set. A set of slide examples with ROI proposals and the actual consensus ROIs are shown in Figure 5.7.

The prediction of the class label of a whole slide image starts with the computation of class probabilities of the proposed feature representations of the associated ROI proposals, following the procedures described in Section 5.4.1. After ROI-level class probabilities of the ROI proposals were predicted by the ROI-level classifier, we used two decision aggregation methods to obtain the slide-level classification results using the ROI-level probabilities of the ROI proposals. We employed the Max-pooling method, with classification threshold set to 0.50 and the frequency threshold set to 0.25, on the ROI-level class probabilities of the ROI proposals to obtain slide-level classification outputs. As an alternative to the Max-pooling based approach, we used the Decision-fusion method to predict slide-level labels by training a classifier on the class frequency histograms from the ROI-level classification results.

We presented the classification performance at slide-level using the proposed ROI-level feature representations, the Penultimate-Feat-Weighted and



(a) Whole slide image (b) ROI proposals (c) Consensus ROIs

Figure 5.7: ROI proposals and the consensus ROIs of the same slide for example whole slide images. The ROI proposals were extracted from the nuclei-dense regions within whole slide images using a sliding window based approach. The automatically extracted ROI proposals of a slide are representative of the regions indicated by the consensus ROIs of the slide.

Table 5.8: The confusion matrix of the slide-level classification results using the Penultimate-Feat-Weighted features for ROI-level predictions and Max-Pooling method for slide-level prediction.

		Predicted			
		Benign	ADH	DCIS	INV
True	Benign	6	10	4	2
	ADH	11	28	7	2
	DCIS	2	7	25	4
	INV	0	1	7	4

Table 5.9: The confusion matrix of the slide level classification results using the Hypercolumn-Feat-Weighted features for ROI-level predictions and Max-Pooling method for slide-level prediction.

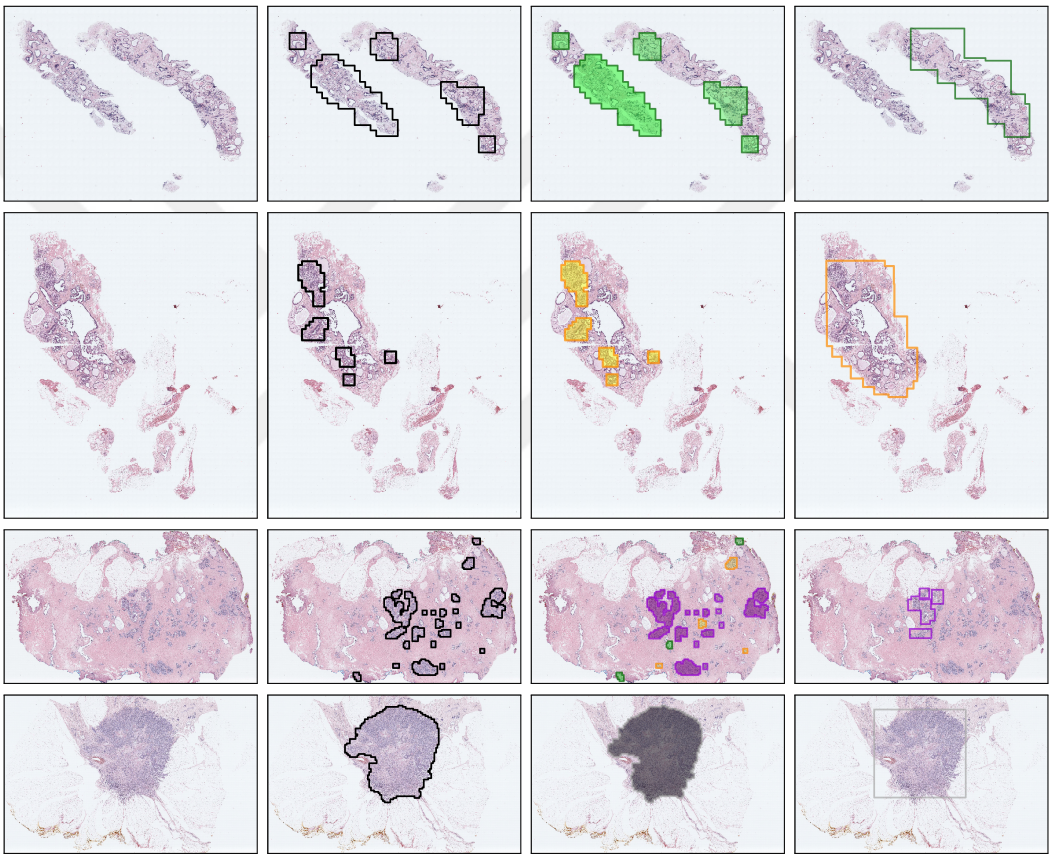
		Predicted			
		Benign	ADH	DCIS	INV
True	Benign	8	9	3	2
	ADH	11	27	8	2
	DCIS	4	6	24	4
	INV	0	2	5	5

Hypercolumn-Feat-Weighted, as well as the slide-level decision aggregation methods, Max-pooling and Decision-fusion, in Table 5.7. The Hypercolumn-Feat-Weighted features had better performance regardless of the selection of the slide-level classification method compared to the Penultimate-Feat-Weighted features. However, the Decision-fusion classifier was not able to predict INV cases and favored diagnoses with more cases present in the training set, such as ADH and DCIS. One of the main reasons is that the Decision-fusion method performs poorly when the number of ROIs vary during the training and the test phases due to the computation of class frequency histograms. Therefore, this method suffers greatly in our slide-level classification setting where the sizes and the number of ROI proposals can vary to a great degree. On the other hand, Max-pooling approach was less sensitive to the ROI imbalance problem, as it was able to correctly classify more cases in minority classes with Benign and INV, while mostly correctly classifying the cases from the majority classes with ADH and DCIS. The confusion matrix of the slide level-classification results using the Penultimate-Feat-Weighted features for ROI-level classification and Max-pooling method for slide-level classification is given in Table 5.8. Similarly, the confusion matrix

of the slide level-classification results using the Hypercolumn-Feat-Weighted features for ROI-level classification and Max-pooling approach for slide-level classification is presented in Table 5.9. The results for Penultimate-Feat-Weighted features and Hypercolumn-Feat-Weighted features combined with Max-pooling method for slide-level classification showed similar performances; the latter could predict more cases with Benign and INV correctly while suffering from a relatively smaller Precision value compared to the former. The ROI-level classification results using the Hypercolumn-Feat-Weighted features were given in Figure 5.8. The consensus ROIs and the associated labels were also provided for comparison. The extracted ROI proposals could cover a large portion of the consensus ROIs and the predictions made by the classifier using the Hypercolumn-Feat-Weighted features matched the diagnostic label of the consensus ROIs. The slide-level labels were computed by the Max-pooling method which was applied on the class probabilities of the ROI-level classifier computed on the ROI-proposals. The max-pooling approach could correctly classify the slide-level labels, as presented in Figure 5.8, regardless of the size or the number of the ROI proposals within the slides. The consensus labels of the slides were Benign, ADH, DCIS and INV from first row to the last row. Please note that the consensus ROIs annotated by the pathologists in the last column of Figure 5.8 do not necessarily cover all of the representations of the corresponding consensus diagnosis in the slide. The consensus ROIs demonstrate a subset of the regions belonging to the most severe diagnosis, i.e. consensus diagnosis, observed in the associated slide.

5.5 Discussion

Convolutional networks operate on fixed-sized patches and make predictions on unseen patches with exactly the same size. The fact that ROIs belonging to the same whole slide image can be drastically different from each other in size, shape and structure, and the fact that the number of ROIs between whole slide images can vary to a great degree make it difficult to analyze ROIs using convolutional networks. This chapter presented a simple yet effective framework to obtain feature representations for variable number of variable sized ROIs in whole slide



(a) Input slide (b) ROI proposals (c) Predictions (d) Consensus

Figure 5.8: ROI-level classification outputs on example slides. From left to right: The input image to the algorithm; the ROI proposals extracted automatically from the input image; the predicted class labels of the ROI proposals using the Hypercolumn-Feat-Weighted features; the annotated ROIs by the pathologists denoting the most severe diagnostically relevant region on the slide (the color of the ROIs denote the label of the ROIs and the associated slide: Benign as green, ADH as yellow, DCIS as purple, INV as gray).

images. The proposed methods operated on the automatically extracted ROI patches. The local structural information within a patch as well as the class probability distribution of the patch from the predictions of the deep convolutional network were preserved in the feature representation of the patch. We used average pooling on these patch-level feature representations to obtain the feature representation of the ROI that the patches were extracted from.

We investigated two methods to extract deep feature vectors for a patch. The first approach involved patch-level penultimate layer activations of the network and the second approach contained features at pixel-level by the hypercolumn convolutional activations to define the feature vector of the patch. In both approaches, the feature vector of the patch was weighted by each class probability score from the network output and the concatenated feature vectors formed the final feature representation of the patch. Finally, the feature representation of the ROI is obtained by average pooling the feature representations of its patches. We demonstrated that the representational power of the proposed approaches outperformed the previous best efforts in extensive quantitative experiments at ROI-level classification of breast histopathology images. We also showed in our experiments that the proposed features could also be successfully used in conjunction with various methods to perform slide-level predictions of the whole slide images. Investigating an end-to-end framework involving deep architectures for both feature representation generation and classification can be considered as a future work of this study.

Chapter 6

Joint Slide-level Multi-class Classification and ROI-level Prediction of Whole Slide Breast Histopathology Images

This chapter introduces a novel weakly supervised learning framework that can model complex relations involving ROI structures, ROI-level latent labels and slide-level diagnoses considering the pathologists' viewing records of whole slide images and their diagnostic observation in the pathology forms. We model ROI-level latent information as well as slide-level diagnostic observations by employing a joint multi-instance multi-label learning approach to simultaneously perform multi-label slide-level classification and ROI-level label inference in whole slide breast histopathology images. As previously discussed in Chapter 4 in much greater detail, the input to a multi-instance learning (MIL) algorithm is a bag which contains variable number of instances with positive and negative examples for the class labels of the bag. If there is at least one instance corresponding to a class label, then the class label of the associated bag is also positive while a class label of the bag is negative if none of the instances is a representative of the class label. In multi-label learning (MLL), the input to the algorithm involves

examples that can be associated with one or more of the class labels. Multi-instance multi-label learning (MIMLL) is the combination of MIL and MLL in the sense that the input is represented by a bag of varying number of instances and is also associated by one or more of the class labels. In each setting, only the bag labels are known during the training; the instance labels are not known a priori. In this study, the bags correspond to whole slide images which are associated with slide-level diagnostic labels whereas the instances come from the candidate ROIs (see Section 3.2 and 6.1.1).

While MIMLL algorithms have been rarely investigated in the field of histopathological image analysis, MIL was investigated in several works for the clinically less significant setting of binary classification [9–13, 64]. MLL of histopathology image analysis in the context of single-instance learning was also studied in some works [14, 15]. Conditional random fields (CRFs) [65, 66] are a set of discriminative models that are naturally part of the MIL domain. However, they have been very scarcely investigated in breast histopathology [16]. Its extension, hidden conditional random fields (HCRFs) [67–69], which provide a more flexible formulation by introducing hidden variables to capture latent information not explicitly observed in the data, have been explored in other domains in computer vision [70, 71]. However, HCRFs have not been explored for the analysis of whole slide images in the weakly supervised setting in any previous related work. In this study, we introduce a HCRF based framework that can simultaneously perform multi-class classification on a whole slide image and make label predictions for each diagnostically relevant region in a weakly supervised setting where only slide-level labels are known during the training and the feature representations of the regions are learned through a deep feature generator network. The overview of this study is presented in Figure 6.1.

The main contribution of this study is twofold. The first contribution is the investigation of MIMLL algorithms on bags of deep feature representations of ROIs, following the network generator approach described in Chapter 5 in a weakly supervised setting, compared to the traditional hand-crafted features based on color, texture and nuclear architecture. We conduct several experiments on both sets of feature representations and make evaluations of the representational capabilities

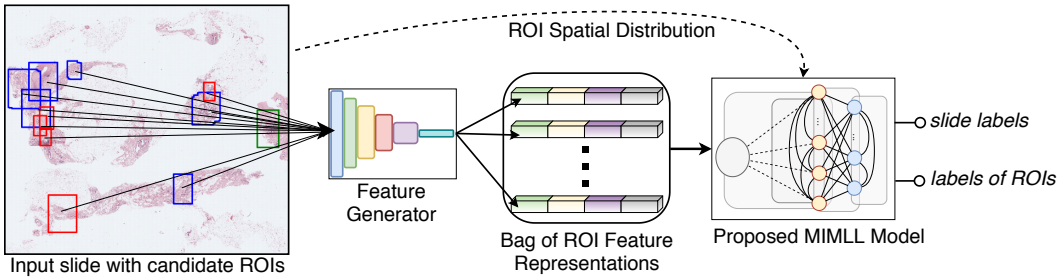


Figure 6.1: The overview of the proposed HCRF based MIMLL approach. The deep feature representation of each ROI in a whole slide image is obtained through a feature generator. The proposed HCRF based MIMLL model simultaneously performs multi-class slide classification and makes label prediction for each ROI on the bag of deep feature representations of ROIs.

in the context of MIMLL. The second and the main contribution of this study is that our model can simultaneously perform multi-class slide-level classification and infer the diagnostic label of each ROI in a whole slide image by considering the spatial distributions of the ROIs, the coherence and the correlations within and between the ROI-level predictions and the slide-level labels. We show that our study outperforms the existing approaches in a weakly supervised setup whereas the competition is also not capable of making ROI-level label predictions directly from slide-level label information.

6.1 Feature Extraction

6.1.1 Elimination of Candidate ROIs

In this study, the diagnostically relevant regions in a whole slide image are automatically discovered by a set of actions defined on the viewing behaviors of the pathologists, as described in Section 3.2. One of the issues of this approach is that it does not consider the risk of detecting the same regions over and over when viewing records of multiple pathologists are combined. Therefore, the redundancy is more prominent when we combine the candidate ROIs of the three experienced pathologists who are likely to visit the most diagnostically relevant regions in slides. Another issue is the erroneous detection of regions either from

background or from areas without nuclei due to human or algorithmic error. Therefore, we employ a set of operations on the candidate ROIs obtained from the three actions on the viewing records of the three experienced pathologists to eliminate the erroneously detected ones that cover mostly the background area or have little to no nuclei, as well as to discard the redundant regions arising from using combined data.

6.1.1.1 Background

A very small number of detected regions have a large ratio of background coverage. We discard a region from the set of candidate ROIs if the foreground coverage of the region, i.e. the ratio of area in foreground to the total area, is below a threshold. We set the foreground threshold to 0.60 based on our empirical observations, and eliminated the regions whose foreground ratio was less than the foreground threshold.

6.1.1.2 Nuclei Density

The regions that have good foreground coverage but contain little to no nuclei lack the descriptive power to use in feature representation. The feature generation network (see Section 6.1.2) extracts patches from nuclei dense areas in the candidate ROIs. We discard a region from the set of candidate ROIs if the area containing nuclei in the region, i.e. the ratio of the area with nuclei to the total area of the region, is below a threshold. In addition, we use another threshold to consider larger regions that have considerable nuclei density, but fall below the area threshold due to the sheer size of the region. We set the area threshold to 0.01 and density threshold to 1000 pixels, at $10\times$ magnification, and eliminated the regions with properties below these thresholds.

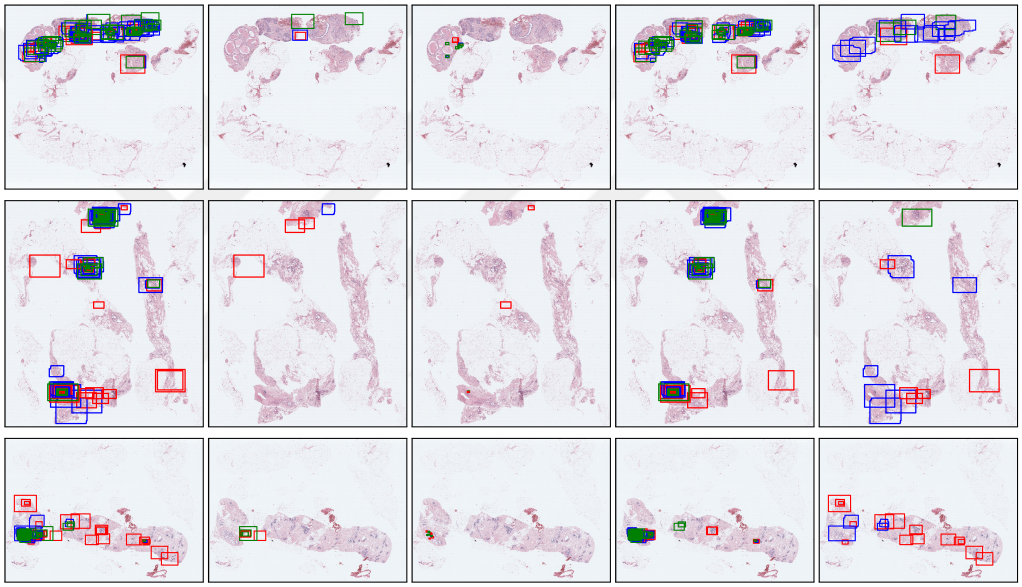
6.1.1.3 Overlap

A large number of detected regions cover small areas which are usually the outcome of zoom-in actions. However, an area in a whole slide image that is populated with several small detected regions may have also been detected by a single large region. We discard these small regions from the set of candidate ROIs if the area covered by them is also covered by a larger region in the same set of candidate ROIs, i.e. the ratio of the intersection of the union of a set of small regions to a larger region in the set of candidate ROIs. The coverage threshold is set to 0.50. Also, we consider a small region to be inside a larger region if their intersection ratio is above an intersection threshold, which we set to 0.75.

The three operations are applied sequentially to remove the redundancy in the candidate ROIs and to weed out the erroneous inputs. A set of example slides with the candidate ROIs from the three actions, discarded candidate ROIs after each subsequent operation and the final candidate ROIs after the operations are presented in Figure 6.2.

6.1.2 ROI-level Deep Feature Generation

Convolutional neural networks (CNNs) have dominated the field of image classification and representation in recent years [27, 72–75]. However, their use-cases are not immediately obvious for the problems in the field of histopathology image analysis [76]. CNNs require their input to be a small fixed-sized image whereas each histopathology slide contains variable number of variable sized ROIs which prevents direct use of the ROIs as inputs. Following the approach presented in Section 5.2, the deep feature representation of an ROI is obtained by a feature generator network using the feature representations of small fixed-sized patches extracted from the nuclei-dense regions in the ROI. Each patch is fed to a fine-tuned deep network and the deep hypercolumn feature vector of a patch is obtained by max-pooling the convolutional channels in select convolutional layers,



(a) (b) (c) (d) (e)

Figure 6.2: The discovery of candidate ROIs in whole slide images include (a) the detection of potential regions with diagnostic relevance from the three actions defined on the viewing records of the pathologists, followed by the identification of (b) regions that cover mostly background area, (c) regions that contain little to no nuclei density, and (d) regions that are covered by a larger region, then finally (e), removing the regions in (b), (c) and (d) from the set of candidate ROIs in (a). Red rectangles denote zoom-in, blue polygons represent slow-panning and green rectangles indicate fixation.

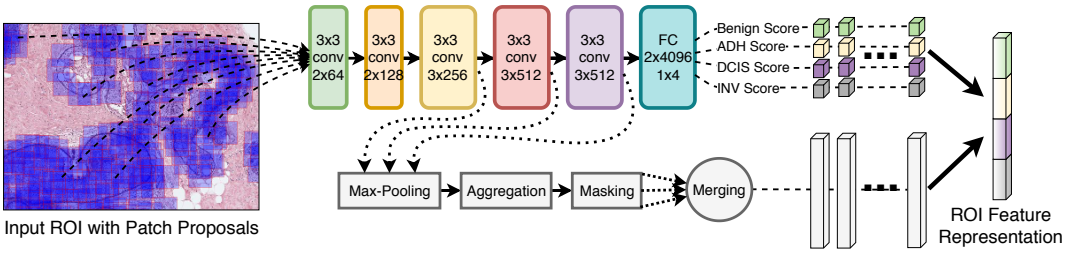


Figure 6.3: The overview of the ROI-level feature generator network. The ROI-level deep feature representations are obtained from the properties of the convolutional channels and network outputs at patch-level.

then aggregating the responses, followed by thresholding the outputs by the median value, and finally merging each vectorized mask. The feature representation of a patch is the aggregation of its deep hypercolumn feature vector with each class specific prediction of the patch from the same network. The deep feature representation of the ROI is obtained by averaging over the feature representations of its patches. This whole process in the form of an ROI-level deep feature generator network is presented in Figure 6.3. The method works regardless of the deep network of choice and regardless of the shape and the size of the ROI. The network is fine-tuned on the consensus ROIs and the consensus labels of a subset of the entire data set which are only used for feature generation purposes. A training/test sample in the learning algorithm (refer to Section 6.2) is a slide from the rest of the data set which is represented by a bag of deep feature representations of the candidate ROIs in the slide and the combined diagnostic labels associated by the pathologists at slide-level. For the rest of the chapter, an ROI will refer to a candidate ROI in a whole slide image unless otherwise stated.

6.2 Learning

In the MIMLL framework, let a data set $\{(X_1, V_1), (X_2, V_2), \ldots, (X_M, V_M)\}$ consist of a collection of samples where M denotes the number of samples. Each sample corresponds to a pair of a bag and a set of labels. A bag X_m contains a set of instances $\{\mathbf{x}_{m1}, \mathbf{x}_{m2}, \ldots, \mathbf{x}_{mn_m}\}$ where n_m is the number of instances in that bag, $\mathbf{x}_{mn} \in \mathbb{R}^d, n = 1, \ldots, n_m$, is the feature vector of the $n'th$ instance of the $m'th$ bag, and d is the number of features. A label set V_m is composed of

class labels $\{v_{m1}, v_{m2}, \ldots, v_{ml_m}\}$ where l_m is the number of labels in that set and $v_{ml} \in \{c_1, c_2, \ldots, c_L\}, l = 1, \ldots, l_m$, is one of L possible class labels.

We adapt the HCRF based MIMLL framework described in [68] for the analysis and the classification of whole slide breast histopathology images. We introduce a membership label vector $\mathbf{y}_m = \{y_{m1}, y_{m2}, \ldots, y_{mL}\}$ where $y_{ml} \in \{-1, +1\}, l = 1, \ldots, L$ with -1 denoting no membership association of the bag X_m with the label c_l and $+1$ denoting a membership association. Even though the relation between a bag X_m and a membership label y_{ml} is known, the relationship between an instance $\mathbf{x}_{mn} \in X_m$ and a membership label y_{ml} is not explicitly provided in the data set. In order to capture such relationships, we introduce a hidden vector $\mathbf{h}_{mn}$ to formulate the association between the instance $\mathbf{x}_{mn}$ and the membership labels $\mathbf{y}_m$. The values of the binary L dimensional vector, $\mathbf{h}_{mn}$, are expected to capture the semantic information between the instance and the bag labels.

For the rest of the chapter, we simplify the notation by dropping the bag indices. The MIMLL problem based on HCRFs can be formulated as the identification of the label vector $\mathbf{y}^*$ given the observation $\mathbf{x}$ as

$$\mathbf{y}^* = \arg\max_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}; \Theta), \tag{6.1}$$

where

$$P(\mathbf{y}|\mathbf{x}; \Theta) = \sum_{\mathbf{h}} P(\mathbf{y}, \mathbf{h}|\mathbf{x}; \Theta) \tag{6.2}$$

$$= \sum_{\mathbf{h}} \frac{1}{Z(\mathbf{x})} \exp\{\Phi(\mathbf{y}, \mathbf{h}, \mathbf{x}; \Theta)\}, \tag{6.3}$$

and $P(\mathbf{y}|\mathbf{x}; \Theta)$ is the posterior distribution of the labels, $\Phi(\mathbf{y}, \mathbf{h}, \mathbf{x}; \Theta)$ is a scale-valued potential function and $Z(\mathbf{x}) = \sum_{\mathbf{y}} \sum_{\mathbf{h}} \exp\{\Phi(\mathbf{y}, \mathbf{h}, \mathbf{x}; \Theta)\}$ is a partition function. We drop Θ from the equations from now on to ease the notation.

6.2.1 Model Definition

The potential function could be decomposed into functions that encode information between bags (slides), bag labels (slide labels), instances (ROIs) and hidden

variables (ROI labels) as

$$\Phi(\mathbf{y}, \mathbf{h}, \mathbf{x}) = \Phi_a(\mathbf{h}, \mathbf{x}) + \Phi_s(\mathbf{h}, \mathbf{x}) + \Phi_{hy}(\mathbf{h}, \mathbf{y}) + \Phi_{yy}(\mathbf{y}), \tag{6.4}$$

where Φ_a models the association between an instance and its label, Φ_s models the spatial relations between the instances in a bag, Φ_{hy} models the consistency of the instance label to the bag label, and Φ_{yy} models the correlations of the bag labels.

Association between an ROI and its diagnostic label: Even though the diagnostic labels of a slide are known, the correspondences of the labels to the ROIs in the slide are not. For a slide, $\Phi_a(\mathbf{h}_i, \mathbf{x}_i)$ is used to model the hidden labels of the ROIs which are not explicitly provided in the data. Unlike slides which can have multiple labels, we assume that each ROI can only be associated with a single diagnostic label. The relation can be modeled by using a local classifier. This potential function, based on the posterior probability $P(\mathbf{h}_i|\mathbf{x}_i; \lambda)$, based on the outputs of the local classifier is then given by

$$\Phi_a(\mathbf{h}, \mathbf{x}) = \sum_i \Phi_a(\mathbf{h}_i, \mathbf{x}_i) \tag{6.5}$$

$$= \sum_i \log P(\mathbf{h}_i|\mathbf{x}_i; \lambda), \tag{6.6}$$

where $1 \leq i \leq n$ and λ denotes the set of the parameters governed by the local classifier.

Spatial relation between the ROIs in a slide: The spatial distribution of the ROIs in a slide contains cues for the ROI labels. We exploit the fact that some diagnostic labels frequently occur in the spatially close regions in a slide. Hence, we propose the following potential function that captures this relation between the pairs of ROIs to improve the predictions of the ROI labels as

$$\Phi_s(\mathbf{h}, \mathbf{x}) = \sum_{p,q} \alpha_{p,q} f_{p,q}(\mathbf{h}, \mathbf{x}) \tag{6.7}$$

$$= \sum_{p,q} \alpha_{p,q} \sum_{i,j} \delta[\![h_{i,p} = 1]\!] \; \delta[\![h_{j,q} = 1]\!] \; \delta[\![\mathbf{x}_i \sim \mathbf{x}_j]\!] \nabla(\mathbf{x}_i, \mathbf{x}_j), \tag{6.8}$$

where $p, q \in \{1, 2, \ldots, L\}$ are label indices, $\alpha_{p,q}$ is the weight parameter, $1 \leq i < j \leq n$, δ is the indicator function whose value is 1 if the predicate is true

and 0 otherwise, $\sim$ denotes a boolean operator that returns true if the pair of ROIs belongs to the same core in the slide and returns false otherwise, $\nabla(\cdot,\cdot)$ is the operation that computes the proximity of a pair of ROIs. The proximity of two ROIs is based on the euclidean distance between them. Each distance is normalized to have a value between 0 and 1. The proximity of a pair of ROIs is computed as the subtraction of their normalized distance from 1.

Coherence between a slide label and the labels of the associated ROIs: A slide is diagnosed as positive for a label if it has at least one region that is associated with that label. This potential function models the coherence between the slide-level label set and the labels of the ROIs. We employ the Ising model [65] to impose consistency between ROI labels $\mathbf{h}$ and slide-level membership labels $\mathbf{y}$ as

$$\Phi_{hy}(\mathbf{h},\mathbf{y}) = \gamma \mathbf{z}^T \mathbf{y}, \tag{6.9}$$

where

$$z_i = \begin{cases} +1, & \text{if } \exists_{1\leq r\leq R}\ h_{r,i} = 1 \\ -1, & \text{if } \forall_{1\leq r\leq R}\ h_{r,i} \neq 1 \end{cases}$$

and $\mathbf{z}$ is an L dimensional binary label vector, γ is the cost that penalizes the inconsistency between $\mathbf{z}$ and $\mathbf{y}$.

Correlations of slide labels: A slide could be associated with one or more of the diagnostic labels. Some diagnostic labels co-occur more frequently than other labels in the pathology forms filled out by the pathologists. Both positively correlated label pairs (frequent co-occurrence) and negatively correlated label pairs (rare co-occurrence) hold valuable information to improve slide-level classification performance. Therefore, we exploit the correlations between pairs of diagnostic labels in the label set of each slide by formulating the potential function as

$$\Phi_{yy}(\mathbf{y}) = \sum_{k,l}\sum_{r,s} \mu_{k,l,r,s} f_{k,l,r,s}(\mathbf{y}) \tag{6.10}$$

$$= \sum_{k,l}\sum_{r,s} \mu_{k,l,r,s} \delta[\![y_k = r]\!] \delta[\![y_l = s]\!], \tag{6.11}$$

where $r, s \in \{-1, +1\}$ are binary (positive and negative) labels, $1 \leq k, l \leq L$ are label indices and $\mu_{k,l,r,s}$ is the weight parameter.

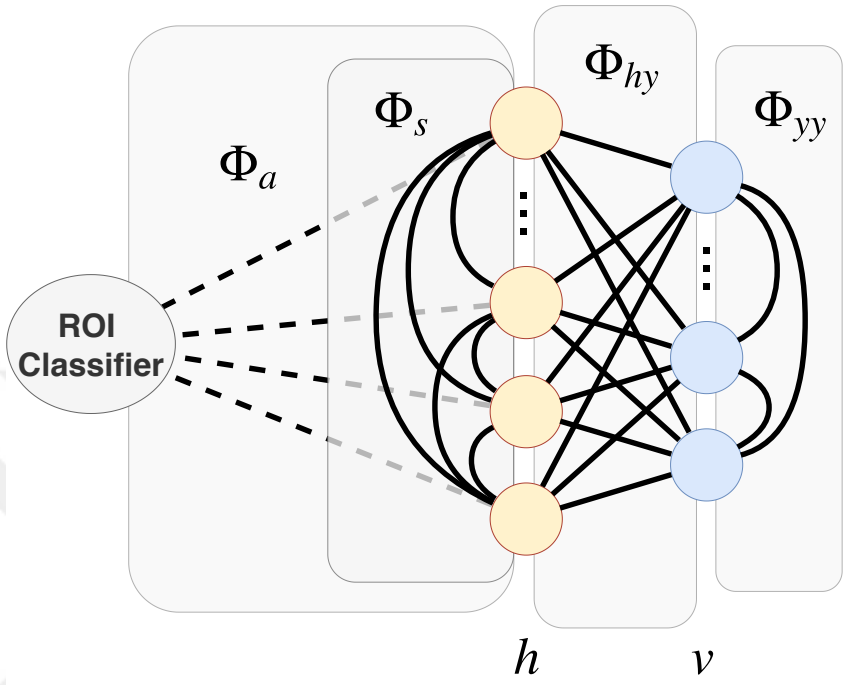


Figure 6.4: The graphical view of the proposed model involving the potential functions; Φ_s, the spatial relation between ROIs in a slide; Φ_{hy}, the coherence between slide labels and the labels of the ROIs in the slide; Φ_{yy}, the correlations of the slide labels; Φ_a, the association between an ROI and its diagnostic label. The proposed model learns the parameters of the potential functions in black straight lines and the parameters of the potential function in dashed lines sequentially, in an alternating optimization scheme.

The combination of the four potential functions defines the complex dependencies, associations and relations both implicitly and explicitly observed in the data. The graph based representation of the proposed HCRF based MIMLL model involving the four potential functions is shown in Figure 6.4.

6.2.2 Training

We estimate the set of parameters $\Theta = \{\alpha_{p,q}, \gamma, \mu_{k,l,r,s}\}$ by maximizing the log-likelihood with respect to the conditional distribution

$$\begin{aligned}\mathcal{L}(\Theta) &= \log \prod_i P(\mathbf{y}|\mathbf{x}_i;\Theta) \\ &= \sum_i \log P(\mathbf{y}|\mathbf{x}_i;\Theta) \\ &=< \log \sum_{\mathbf{h}} P(\mathbf{y},\mathbf{h}|\mathbf{x};\Theta) >_{\tilde{\Gamma}},\end{aligned} \tag{6.12}$$

where $\tilde{\Gamma}$ denotes the empirical distribution and $< \cdot >_{\tilde{\Gamma}}$ corresponds to the expectation with respect to $\tilde{\Gamma}$. Such estimation procedures are known to have problems when it comes to generalization. Hence, we introduce a penalization term to improve its generalization capabilities using the log of a Gaussian prior with variance σ^2 as

$$\mathcal{L}(\Theta) =< \log \sum_{\mathbf{h}} P(\mathbf{y},\mathbf{h}|\mathbf{x};\Theta) >_{\tilde{\Gamma}} - \frac{1}{2\sigma^2}\|\Theta\|^2, \tag{6.13}$$

with $p(\Theta) \sim \exp\left(-\frac{1}{2\sigma^2}\|\Theta\|^2\right)$. Given the hidden variables in the optimization formulation, we resort to using the well-known *Expectation Maximization* (EM) algorithm [77] to solve this optimization problem. In the Expectation step (E-step), we can write the Q function that we want to maximize using the expectation of $\mathcal{L}(\Theta)$ under the parameter estimates at step t as

$$Q(\Theta,\Theta^{(t)}) =< \mathbf{E}_{\mathbf{h}|\mathbf{y},\mathbf{x};\Theta^{(t)}} \log \sum_{\mathbf{h}} P(\mathbf{y},\mathbf{h}|\mathbf{x};\Theta) >_{\tilde{\Gamma}} - \frac{1}{2\sigma^2}\|\Theta\|^2, \tag{6.14}$$

where $\mathbf{E}_{\mathbf{h}|\mathbf{y},\mathbf{x};\Theta^{(t)}}$ denotes the expectation given the current estimated conditional probability $P(\mathbf{h}|\mathbf{y},\mathbf{x};\Theta)$. Then, we maximize the Q function in the M-step to obtain the new parameters Θ^{t+1} as

$$\Theta^{(t+1)} = \arg\max_{\Theta} Q(\Theta,\Theta^{(t)}). \tag{6.15}$$

In order to maximize the Q function, we take its derivatives with respect to the parameters $\alpha_{p,q}, \gamma$, and $\mu_{k,l,r,s}$ as

$$\frac{\partial Q}{\partial \alpha_{p,q}} =< \mathbf{E}_{\mathbf{h}|\mathbf{y},\mathbf{x};\Theta^{(t)}} f_{p,q}(\mathbf{h},\mathbf{x}) >_{\tilde{\Gamma}} - < f_{p,q}(\mathbf{h},\mathbf{x}) >_{\Gamma} - \frac{1}{\sigma^2}\alpha_{p,q}, \tag{6.16}$$

$$\frac{\partial Q}{\partial \gamma} =< \mathbf{E}_{\mathbf{h}|\mathbf{y},\mathbf{x};\Theta^{(t)}} \mathbf{z}^T\mathbf{y} >_{\tilde{\Gamma}} - < \mathbf{z}^T\mathbf{y} >_{\Gamma} - \frac{1}{\sigma^2}\gamma, \tag{6.17}$$

$$\frac{\partial Q}{\partial \mu_{k,l,r,s}} =< \mathbf{E}_{\mathbf{h}|\mathbf{y},\mathbf{x};\Theta^{(t)}} f_{k,l,r,s}(\mathbf{y}) >_{\tilde{\Gamma}} - < f_{k,l,r,s}(\mathbf{y}) >_{\Gamma} - \frac{1}{\sigma^2}\mu_{k,l,r,s}, \tag{6.18}$$

where Γ denotes the model distribution. All parameter update equations require the computation of the expectation over data distribution of some statistics; $< \cdot >_{\tilde{\Gamma}}$, and the computation of expectation over the model distribution; $< \cdot >_{\Gamma}$. The partition function requires summing over all possible values of the image label and over all possible values that each ROI can take in each slide. Hence, computing the derivatives which include expectations under the model distribution is not feasible due to the existence of the partition function in the denominator. We can resort to approximate inference algorithms such as the sampling based method, Markov chain Monte Carlo (MCMC) [78]. However, Markov chains suffer from very slow convergence rates, and it is often not easy to decide when to stop the chain to start the sampling. Therefore, we employ the *Contrastive Divergence* (CD) algorithm to speed up the convergence process which was utilized for efficient training of the family of Products of Experts models [79, 80].

Given a system with visible and hidden units, the initial ($t = 0$) step of the CD algorithm starts with clamping data into visible units and updating the hidden units, and one step of CD (CD-1) is completed by reconstructing the visible units, followed by updating the hidden units. CD-k algorithm is basically repeating this update chain of visible and hidden units k times. When $k \to \infty$, the statistics of the expectation over the model distribution can be replaced by the samples from the visible and the hidden units at the end of the chain, and similarly, the statistics of the expectation over the data distribution is replaced by the data clamped to the visible units and by the samples from the hidden units at the start of the chain. A simple illustration of the CD algorithm is provided in Figure 6.5. These iterations are time-consuming as the system might take a

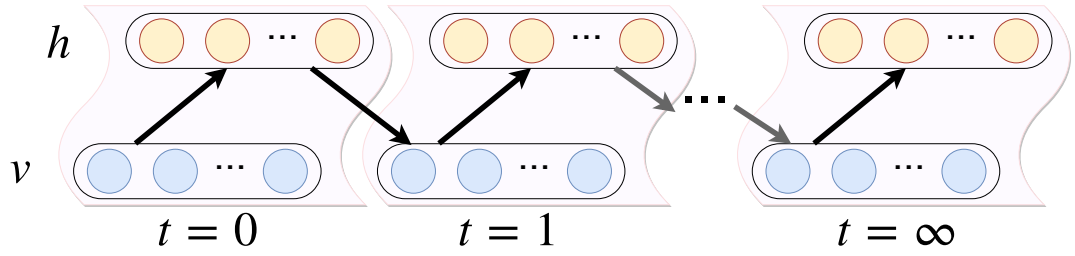


Figure 6.5: The visual representation of the Contrastive Divergence (CD) algorithm. The algorithm starts with clamping data points to the visible units, followed by sampling from the hidden units, and repeating these steps until the algorithm reaches equilibrium when the values of the visible and the hidden units do not change.

very long time to reach equilibrium. However, in some models (i.e. Restricted Boltzmann Machines), CD-1 was found to be performing surprisingly well even tough this approach does not follow the gradient of the loglikelihood function directly. Instead of waiting for the system to reach equilibrium, we stop the chain after CD-1 is completed. As before, the statistics of the expectation over the model distribution are then replaced by the samples from the visible and the hidden units at the end of the chain, and the statistics of the expectation over the data distribution are replaced by the data clamped to the visible units and by the samples from the hidden units at the start of the chain.

In our study, we are incorporating the idea of CD-1 to compute the statistics in the learning rules due to the running time of the algorithm becoming infeasible with increasing number of cycles. In our model, the slide labels represent the states of the visible units, whereas the ROI labels represent the states of the hidden units. First, we clamp the slide labels to the visible units ($\mathbf{y}^0 = \mathbf{y}$) where $\mathbf{y}^0$ denotes the state of the visible units at CD-0, and sample from the hidden units, $\mathbf{h}^0 \sim p(\mathbf{h}|\mathbf{y} = \mathbf{y}^0, \mathbf{x}; \Theta)$. This step is called the positive/clamped phase. We complete CD-1 by sampling from the visible units, $\mathbf{y}^1 \sim p(\mathbf{y}|\mathbf{h} = \mathbf{h}^0, \mathbf{x}; \Theta)$, followed by sampling from the hidden units $\mathbf{h}^1 \sim p(\mathbf{h}|\mathbf{y} = \mathbf{y}^1, \mathbf{x}; \Theta)$ which is called the negative/free phase. Note that, while we use the slide labels as the states of the visible units in the clamped phase, we sample from the visible units in the free phase. Due to the associations defined between visible units, block sampling from the visible units is not possible. We sample from each visible unit

one by one by Gibbs sampling by computing the probability of a visible unit, $y_i, i \in \{1, \ldots, L\}$, being turned on as

$$p(y_i = +1|\mathbf{y}_{\bar{i}}, \mathbf{h}, \mathbf{x}; \Theta) \tag{6.19}$$

$$= \frac{p(y_i = +1, \mathbf{y}_{\bar{i}}, \mathbf{h}|\mathbf{x}; \Theta)}{p(\mathbf{y}_{\bar{i}}, \mathbf{h}|\mathbf{x}; \Theta)} \tag{6.20}$$

$$= \frac{p(y_i = +1, \mathbf{y}_{\bar{i}}, \mathbf{h}|\mathbf{x}; \Theta)}{\sum_{y_i} p(y_i, \mathbf{y}_{\bar{i}}, \mathbf{h}|\mathbf{x}; \Theta)} \tag{6.21}$$

$$= \frac{p(y_i = +1, \mathbf{y}_{\bar{i}}, \mathbf{h}|\mathbf{x}; \Theta)}{p(y_i = +1, \mathbf{y}_{\bar{i}}, \mathbf{h}|\mathbf{x}; \Theta) + p(y_i = -1, \mathbf{y}_{\bar{i}}, \mathbf{h}|\mathbf{x}; \Theta)} \tag{6.22}$$

$$= \frac{\exp\{\Phi(\mathbf{y}^+, \mathbf{h}, \mathbf{x}; \Theta)\}/Z(\mathbf{x})}{(\exp\{\Phi(\mathbf{y}^+, \mathbf{h}, \mathbf{x}; \Theta)\} + \exp\{\Phi(\mathbf{y}^-, \mathbf{h}, \mathbf{x}; \Theta)\})/Z(\mathbf{x})} \tag{6.23}$$

$$= \frac{\exp\{\Phi_{hy}(\mathbf{h}, \mathbf{y}^+) + \Phi_{yy}(\mathbf{y}^+)\}}{\exp\{\Phi_{hy}(\mathbf{h}, \mathbf{y}^+) + \Phi_{yy}(\mathbf{y}^+)\} + \exp\{\Phi_{hy}(\mathbf{h}, \mathbf{y}^-) + \Phi_{yy}(\mathbf{y}^-)\}} \tag{6.24}$$

where $\mathbf{y}^+ = \{y_i = +1, \mathbf{y}_{\bar{i}}\}, \mathbf{y}^- = \{y_i = -1, \mathbf{y}_{\bar{i}}\}, \bar{i} = \{1, \ldots, L\} \setminus i$. As expected, the probability of a state of a visible unit (one of the class labels of a slide) being turned on or off is determined by the potentials that involve the slide labels.

The steps of sampling from the hidden units are similar but not identical to the steps of sampling from the visible units. As previously stated, ROIs can only be associated with a single class label, which enforces that a hidden unit can only take on a single class value. Given a set of hidden units, $\mathbf{h} = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n\}$, we compute the ratio of the probability of a hidden unit having a new random class label q, $h_{kq} = +1, q \in \{1, \ldots, L\}, \mathbf{h}_{k\bar{q}} = -\mathbf{1}, \bar{q} = \{1, \ldots, L\} \setminus q$, to its current configuration with class label p, $h_{kp} = +1, p \in \{1, \ldots, L\}, \mathbf{h}_{k\bar{p}} = -\mathbf{1}, \bar{p} = \{1, \ldots, L\} \setminus p$, as

$$\tau = \frac{p(h_{kq} = +1, \mathbf{h}_{k\bar{q}} = -\mathbf{1}|\mathbf{h}_l, \mathbf{y}, \mathbf{x}; \Theta)}{p(h_{kp} = +1, \mathbf{h}_{k\bar{p}} = -\mathbf{1}|\mathbf{h}_l, \mathbf{y}, \mathbf{x}; \Theta)} \tag{6.25}$$

$$= \frac{p(h_{kq} = +1, \mathbf{h}_{k\bar{q}} = -\mathbf{1}, \mathbf{h}_l, \mathbf{y}|\mathbf{x}; \Theta)/\sum_{\mathbf{h}_k} p(\mathbf{h}_k, \mathbf{h}_l, \mathbf{y}|\mathbf{x}; \Theta)}{p(h_{kp} = +1, \mathbf{h}_{k\bar{p}} = -\mathbf{1}, \mathbf{h}_l, \mathbf{y}|\mathbf{x}; \Theta)/\sum_{\mathbf{h}_k} p(\mathbf{h}_k, \mathbf{h}_l, \mathbf{y}|\mathbf{x}; \Theta)} \tag{6.26}$$

$$= \frac{\exp\{\Phi(\mathbf{y}, \mathbf{h}_k^q, \mathbf{h}_l, \mathbf{x}; \Theta)\}/Z(\mathbf{x})}{\exp\{\Phi(\mathbf{y}, \mathbf{h}_k^p, \mathbf{h}_l, \mathbf{x}; \Theta)\}/Z(\mathbf{x})} \tag{6.27}$$

$$= \frac{\exp\{\Phi_a(\mathbf{h}_k^q, \mathbf{h}_l, \mathbf{x}) + \Phi_s(\mathbf{h}_k^q, \mathbf{h}_l, \mathbf{x}) + \Phi_{hy}(\mathbf{h}_k^q, \mathbf{h}_l, \mathbf{y})\}}{\exp\{\Phi_a(\mathbf{h}_k^p, \mathbf{h}_l, \mathbf{x}) + \Phi_s(\mathbf{h}_k^p, \mathbf{h}_l, \mathbf{x}) + \Phi_{hy}(\mathbf{h}_k^p, \mathbf{h}_l, \mathbf{y})\}} \tag{6.28}$$

where $\mathbf{h}_l = \mathbf{h} \setminus \mathbf{h}_k$ and $\mathbf{h}_k^q = \{h_{kq} = +1, \mathbf{h}_{k\bar{q}} = -\mathbf{1}\}, \mathbf{h}_k^p = \{h_{kp} = +1, \mathbf{h}_{k\bar{p}} = -\mathbf{1}\}$. Consequently, we accept the new state, $\mathbf{h}_k^q$ with probability $min(1, \tau)$ and update

the state of the hidden unit accordingly. The sampling procedure is repeated for each hidden unit. As observed from the probabilities, the state of the hidden unit $\mathbf{h}_k$ depends on the states of the visible units $\mathbf{y}$ as well as the states of the other hidden units $\mathbf{h}_l$. Similarly, the state of the visible unit y_i depends on the states of the set of hidden units $\mathbf{h}$ as well as the states of the other visible units $\mathbf{y}_{\bar{i}}$. In other words, when the CD algorithm starts, after clamping the image label to the visible unit, sampling from a hidden unit involves the computation of the potential functions that are evaluated at the current states of the visible units and at the current states of the other hidden units. We can determine the initial states of the hidden units randomly but we can also make informed guesses for the initial states of the hidden units based on the label information at slide level. First, we initialize the labels of the ROIs in a slide with the corresponding slide-level label. If the slide has multiple positive classes, we also associate each ROI with every positive class separately and add the resulting ROI-class pair in the ROI-training set. After populating the ROI-training set from every bag, we train a shallow Neural Network on the deep feature representations of the ROIs and their labels. The network is then used to assign a single class label to each ROI in every slide. We use these labels as the initial labels of the ROIs (and the initial states of the hidden units) in our learning algorithm. The learning algorithm employs EM algorithm for sequential optimization of the model parameters. First, we train the local classifier on the deep feature representations of the ROIs and labels, over all slides. We use a Neural Network, denoted by the weights, λ, with sigmoid hidden units which output one of the L class labels with a softmax activation function. Then, fixing the weights of the network, we update the parameters, $\alpha_{p,q}, \gamma, \mu_{k,l,r,s}$ from the statistics computed from the CD-1 algorithm. The statistics in the expectation over data distribution come from the samples obtained in the clamped phase and the statistics in the expectation over the model distribution are computed from the samples in the free phase. The statistics for both phases are computed for each slide separately, and the average statistics over the slides are used in the update equations of the model parameters (6.16)-(6.18). Finally, we update the ROI labels as the set of labels that maximizes the conditional probability $P(\mathbf{h}|\mathbf{y}, \mathbf{x}; \Theta)$ as

$$\mathbf{h}^* = \arg\max_{\mathbf{h}} P(\mathbf{h}|\mathbf{y}, \mathbf{x}; \Theta). \tag{6.29}$$

The training algorithm stops when we reach a pre-determined number of epochs or when the change in the parameters becomes negligible. The inference of the optimal label of a new image from the posterior distribution is done using the *Maximum Posterior Marginal* (MPM) [65, 81]. As the computation of MPM involves marginalization over a very large number of variables, we use the approximate inference method, Gibbs sampling [78] to speed up the convergence. A similar procedure is applied for the inference of the ROI label, **h**.

6.3 Classification

We conducted ROI-level and slide-level experiments to evaluate the performance of the proposed HCRF based MIMLL model, which from now on we refer as MimlHCRF. Training and test contained bags of instances in the form of deep feature representations which were compared to the hand-crafted color, texture and nuclear features. Each bag was associated with the combined slide-level label sets from the pathology reports of the three experienced pathologists. We also utilized the consensus labels for the ROI-level evaluations in our experiments.

6.3.1 Slide-level Classification

The output of the proposed approach was the single-label prediction of each ROI in the slide and the multi-label prediction of labels of the slide. Since the existing MIMLL algorithms provided only slide-level predictions and were not capable of doing ROI-level inference simultaneously, we compared only the slide-level performance of the proposed model to other MIMLL approaches; MimlSVM, MimlNN, MimlSVMmi and M^{3}Miml, all of which were previously tested extensively on the slide-level classification task in whole slide breast histopathology images [6]. The slide-level classification experiments also involved the evaluation of various combinations of the potential functions. Please note that the experiments included the ROIs that potentially corresponded to some of the diagnostically relevant regions in the slides which were identified from the actions

defined on the viewing records of the pathologists as described in Section 6.1.1. Such regions in whole slide images can also be automatically identified in case no such information is available at test time [38, 39, 82, 83].

6.3.2 ROI-level Classification

MimlHCRF model simultaneously predicted the labels of the ROIs while performing multi-class classification of the slides. Each ROI was assigned a single diagnostic label by the model whose input during the training involved one or more of the slide-level diagnostic labels, i. e. the model had to make educated ROI-level guesses from very coarse slide-level information. Due to the variable number of ROIs and the limited number of consensus ROIs, ROI-level classification included the comparison of the most severe ROI-level prediction to the most severe diagnosis in the slide.

6.4 Experimental Results

The data set used in this study was described in Section 3.1 in great detail. In our experiments, we used a subset of the data set involving the consensus ROIs and the associated consensus labels to fine-tune a deep convolutional network for ROI-level feature generation. The MimlHCRF model used a separate subset of the data set that involved only the viewing records of the pathologists, i.e. a set of bags of deep feature representations of the ROIs, and the individual interpretations of the pathologists of each slide as a combined label set of the pathologists.

6.4.1 Experimental Setting

The data set of 240 slides was split into three sets: 60 slides were used for fine-tuning the deep convolutional network, 120 slides were used for training the

Table 6.1: The distribution of the most severe diagnostic categories in the data set. 60 slides were used for fine-tuning a deep network for learning feature representations, 120 slides were used for training the proposed model, and 60 slides were used for test.

	Feature Generation	Training	Test	Total
Benign without atypia	16	22	18	56
Atypical ductal hyperplasia	17	48	18	83
Ductal carcinoma in situ	22	38	19	79
Invasive cancer	5	12	5	22
Total	60	120	60	240

Table 6.2: Statistics of class combinations in the training and test data. For every combination of diagnostic labels (multi-label setting), the total number of slides and the average number of ROIs discovered from the viewing behavior of the pathologists per slide are shown for the training and the test sets.

	Number of slides		Average number of ROIs per slide	
	Training	Test	Training	Test
Benign	23	18	15.1	13.2
ADH	3	0	4.7	-
DCIS	17	8	12.8	11.4
INV	4	3	8.5	17.7
Benign+ADH	44	18	15.2	12.8
Benign+DCIS	10	6	14.2	14.0
Benign+INV	6	3	11.5	17.7
ADH+DCIS	0	0	-	-
ADH+INV	0	0	-	-
DCIS+INV	5	1	14.2	15.0
Benign+ADH+DCIS	5	1	23.2	17.0
Benign+ADH+INV	0	0	-	-
Benign+DCIS+INV	3	1	15.7	11.0
ADH+DCIS+INV	0	0	-	-
Benign+ADH+DCIS+INV	0	0	-	-

MimlHCRF model, and the remaining 60 slides were used for test as presented in Table 6.1. A whole slide image in the data set was represented by the bag of deep feature representations of the combined ROIs. In addition, each slide was associated with a slide-level combined label set collected from the pathology reports of the pathologists. A slide, on average, contained 13.58 ± 7.81 ROIs after the post-processing steps as described in Section 6.1.1 and it was associated with, on average, 1.70 ± 0.66 labels for the 4-class setting, involving the classes; benign without atypia (Benign), atypical ductal hyperplasia (ADH), ductal carcinoma in situ (DCIS) and invasive cancer (INV). A more detailed breakdown of the number of ROIs for every combination of slide-level combined label sets in the training and test sets are shown in Table 6.2. Each slide also had a single consensus label that was assigned jointly by the pathologists which corresponded to the most severe diagnosis observed in the slide.

Our first experiment involved the comparison of the traditional feature representations of color, texture and nuclear architectural features following the procedure in Section 4.1 with the deep feature representations. We selected the VGG16 architecture [59] as the backbone to our ROI-level deep feature generator network; more details on this were presented in Section 5.2. We selected the VGG16 architecture due to its relatively large depth and representational capabilities. However, the generator network can utilize any off-the-shelf convolutional architecture as its backbone. The details of the VGG16 architecture that we utilized in our experiments as the backbone of the deep feature generator network are as follows. We employed batch normalization before the ReLU activations. The generalization performance of the network was improved by the augmentation techniques applied on the patches such as random rotation, random horizontal and vertical flipping, and random perturbations on the Hue channel of the input image. We fine-tuned the network on patches with 224×224 pixels which were extracted from the consensus ROIs of the whole slide image which were part of the 60 separate histopathology images. The loss function of the network was the cross-entropy loss, and the training involved batches of 8 patches. In addition, we employed Adam optimizer with a learning rate set to $1e{-}4$ and it was set to

decay exponentially with every epoch, with an initial value of 0.01. We also employed dropout by randomly turning off 75% of the fully connected layers during training. The VGG16 network had 13 convolutional layers which we denote as $conv-\{1_1, 1_2, 2_1, 2_2, 3_1, 3_2, 3_3, 4_1, 4_2, 4_3, 5_1, 5_2, 5_3\}$ and three fully-connected layers which we denote as $fc-\{1, 2, 3\}$. In our experiments, we utilized the convolutional activations in $conv-\{3_3, 4_3, 5_3\}$ to obtain the deep feature representations of the ROIs. The experiments involved 120 slides for training and validation, and 60 slides for test. We used a shallow neural network with one hidden layer of 20 neurons as the ROI-level local classifier in the potential function Φ_a. We initialized the parameter γ to 2.5 to ensure consistency between ROI predictions and the combined label sets of the pathologists while the system was unsteady since the only reliable and available information in the start of the training was the slide-level labels. The large initial number of γ lets the system make decent ROI-level predictions from the available slide-level information. Even though the large value cannot guarantee the correct label identifications of the ROIs, it can guarantee the similarity of the label distribution of the ROIs to the slide-level labels. We also set $\alpha_{p,q}$ as well as $\mu_{k,l,r,s}$ to 0. The values of the parameters when the system reached to a steady configuration after several epochs of training are presented in Figure 6.6. For example, Figure 6.6 (d) shows that there is a weak negative co-occurrence between ADH and Benign, meaning that if the slide does not have ADH, the model cannot identify a positive Benign presence. Similarly, Figure 6.6 (d) also shows that if the slide does not have INV, the model identifies a strong presence of ADH in the slide.

6.4.2 Evaluation Criteria

We assessed the quantitative performance of the MIMLL classifiers by comparing the predicted labels to the combined label set assigned by the three pathologists. The training and the test sets came from different splits, as shown in Table 6.1, where each slide belonged to a different patient. Given a test set with N samples $\{(\mathcal{X}_n, \mathcal{Y}_n)\}_{n=1}^{N}$, $\mathcal{Y}_n$ denoting the target labels for the $n'th$ sample, let $f(\mathcal{X}_n)$ be a function that corresponded to a classifier evaluating $\mathcal{X}_n$ and returning

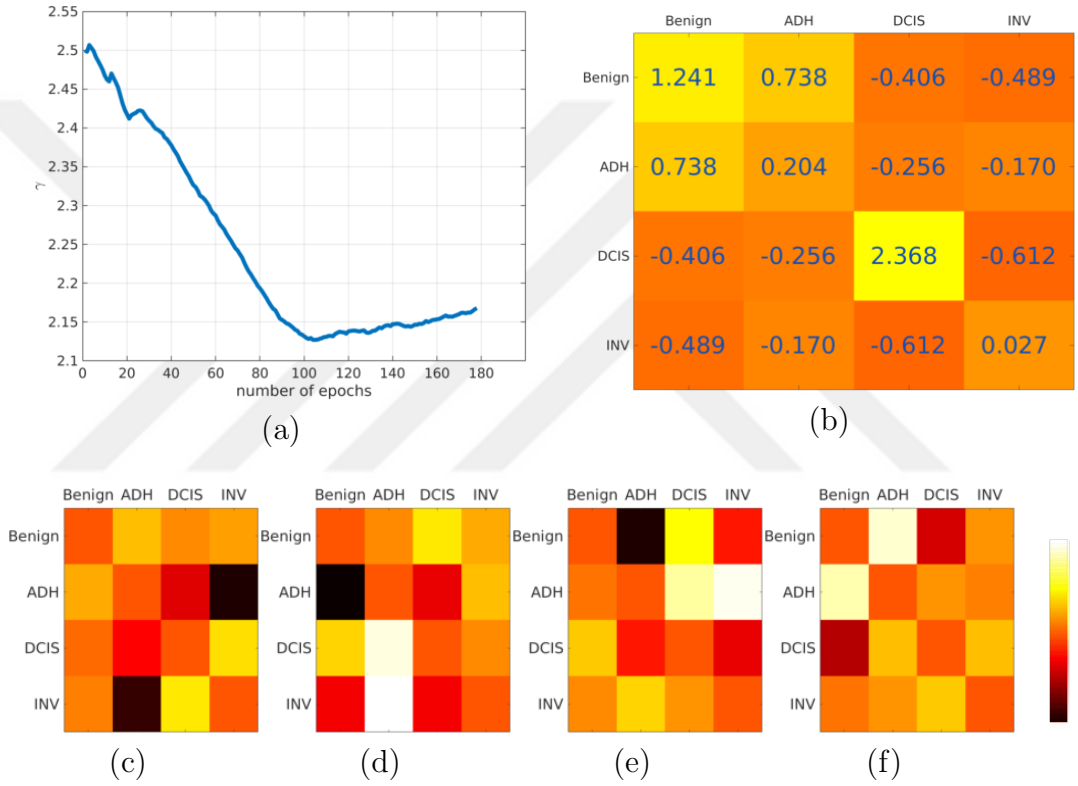


Figure 6.6: The converged values of the parameters of the model. (a) The change in γ parameter through the training epochs represents the dependence of the system on the parameter. (b) The converged values of $\alpha_{p,q}$ for every pair of p, q values from the set of 4 classes and (c-f) the converged values of $\boldsymbol{\mu}_{k,l}$ were from when the system least depended on the γ parameter for every pair of k, l values from the set of 4 classes. Negative co-occurrence (c), negative-positive occurrence (d), positive-negative occurrence (e) and positive co-occurrence (f) are shown in colors from black to red, yellow and white to represent increasing values for the parameters.

the predicted set of class labels. From the multi-label classification literature [84, 85], we selected the most commonly used evaluation criteria that were applicable for our study where the predictions corresponded to the sets of discrete-valued class labels:

- $hammingLoss(f) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L}|f(\mathcal{X}_n)\triangle\mathcal{Y}_n|$, where $\triangle$ operator computes the symmetric distance between two sets. This criterion evaluates the fraction of wrong labels (i.e. false positives or false negatives) to the total number of labels.
- $jaccardIndex(f) = \frac{1}{N}\sum_{n=1}^{N}\frac{|f(\mathcal{X}_n)\cap\mathcal{Y}_n|}{|f(\mathcal{X}_n)\cup\mathcal{Y}_n|}$. Jaccard index, also called accuracy, calculates the fraction of the number of correctly classified labels to the number of labels in the union of predicted and true labels.
- $precision(f) = \frac{1}{N}\sum_{n=1}^{N}\frac{|f(\mathcal{X}_n)\cap\mathcal{Y}_n|}{|f(\mathcal{X}_n)|}$. Precision is the fraction of the correctly classified labels to the number of predicted labels.
- $recall(f) = \frac{1}{N}\sum_{n=1}^{N}\frac{|f(\mathcal{X}_n)\cap\mathcal{Y}_n|}{|\mathcal{Y}_n|}$. Recall is the fraction of the correctly classified labels to the number of true labels.
- $f\text{-}measure(f) = \frac{2\cdot precision(f)\cdot recall(f)}{precision(f)+recall(f)}$. F-measure is the *harmonic mean* of precision and recall.

We note that the performance of an algorithm is better when its evaluated Hamming loss is small, and Jaccard index, precision, recall as well f-measure values are large.

6.4.3 Slide-level Classification Results

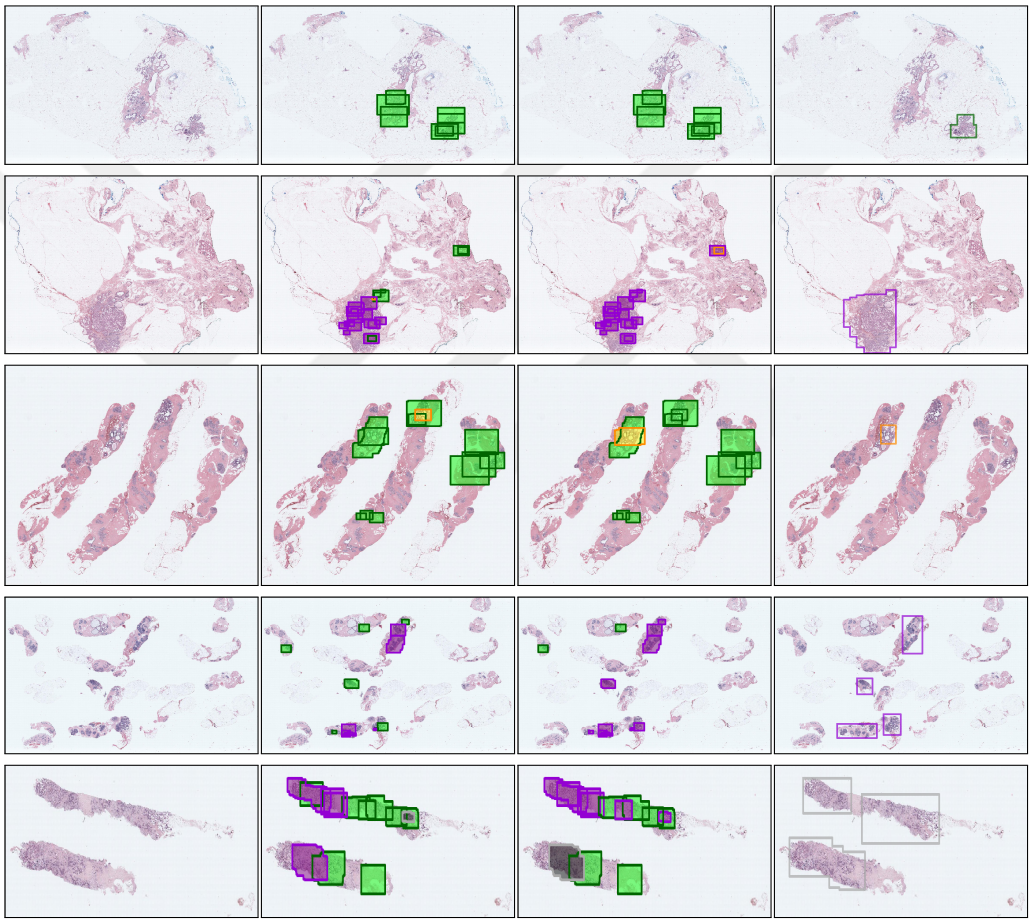
The experimental results of the baseline MIMLL models and the proposed MIML-HCRF model trained on the bags of traditional hand-crafted features of ROIs, and on the bags containing ROI-level feature representations from the deep generator network are presented in Table 6.3. The correct label of a bag included the combined slide-level label sets of the pathologists. The classifiers which were

trained on the deep feature sets had overall better performance compared to the classifiers trained on the traditional feature sets. Among the baseline models, MimlNN performed the best in four out of the five evaluation criteria while MimlSvm had the best precision value. More notably, the best results were obtained when the feature sets included the deep feature representations. In all settings, the classifiers trained on bags of deep feature representations surpassed their counterparts which were trained on bags of hand-crafted traditional features. Please note that the classifier in M^3Miml model which was trained on traditional features performed especially poorly despite our efforts to improve its performance by grid search for hyperparameters and better initialization. We believe that this was due to the small number of slides in the training set which went down from 180 to 120 compared to the experiments in Chapter 4, as well as the smaller capacity of the traditional features for the weakly supervised learning setting when there were not sufficient data available. The same algorithm when trained on the deep feature representations performed much better. When we investigated the results of the proposed model, the comparison of the baseline MIMLL methods with the MimlHCRF model demonstrated the superiority of the proposed approach in slide-level classification, regardless of the selection of feature representations. Similar to the baseline methods, the performance of MimlHCRF was the best in every evaluation criterion when the proposed model was trained on the bags of deep feature representations. The model achieved the smallest Hamming loss, 0.191, the best accuracy, 0.665, precision, 0.815, recall, 0.788 and f-measure, 0.754. The second-best performing model was the baseline MimlNN model, with a value of 0.220 for Hamming loss and 0.692 for f-measure. The rest of the competition was far behind in every evaluation criterion. Similarly, the MimlHCRF model was the best performing slide-level classifier when the features included the traditional hand-crafted features. However, the algorithm made more false positive predictions compared to the next-best performing model, MimlNN, as seen from the relatively larger Hamming loss and smaller precision value. The experimental results indicated that the proposed algorithm made comparatively more false claims than the best performing competition but it was overall the better classifier netting the best accuracy, recall and f-measure values despite being bottle-necked by the traditional features.

We also experimented on different subsets of the potential functions to investigate the performance of different combinations. Table 6.3 presents the results regarding these experiments. Models involving different combinations of the potential functions were trained on the bags of deep feature sets of ROIs using the combined slide-level label sets of the pathologists, in the same fashion as the previous experiment. MimlHCRF$_{1,2,3}$ corresponded to the model when we utilized the potential functions $\Phi_a, \Phi_s, \Phi_{hy}$. The model did not consider the slide-level label co-occurrence which degraded its slide-level classification performance. The following two models, MimlHCRF$_{1,3}$ and MimlHCRF$_{1,2}$, only utilized the potential functions of Φ_a, Φ_{hy} and Φ_a, Φ_s, respectively. Therefore, both models barely converged and as a result, they performed very poorly. The high recall values were observed mostly because the models made too many positive class predictions, therefore lowering the precision values. MimlHCRF$_{1,2}$ especially suffered from not exploiting the only reliable information during training: the slide-level labels. In the light of these results, we see that every potential function complements one another and each plays a different pivotal role in the formulation of MimlHCRF.

6.4.4 ROI-level Classification Results

Our model, MimlHCRF, was capable of inferring ROI labels during slide-level multi-class classification. The ROI-level labels inferred by the model on several example whole slide images are shown in Figure 6.7. We also presented the predictions of the ROIs from the local classifier. It was used as the building block of the potential function that defined the association between an ROI and its label (Φ_a). In addition, we included the consensus ROIs with their associated labels from the same whole slide images as reference. The outputs suggest that the MimlHCRF model was able to correct the initial predictions of the local classifier by exploiting the information from the rest of the potential functions. Please note that, ROI-level classification was a very challenging task for the model as it was only trained using the slide-level combined label sets and the initial training of the local classifier also involved the same information.



(a) Whole slide image (b) Local classifier (c) MimlHCRF (d) Consensus ROIs

Figure 6.7: Example whole slide images, ROI-level predictions by the local classifier and by the MimlHCRF model are shown next to the consensus ROIs colored with the corresponding consensus labels of the same slides for comparison. The predictions of the model generally captured the diagnostic nature of the regions denoted by the colored consensus ROIs. The four colors represent the four diagnostic labels; Benign as green, ADH as yellow with orange border for better viewing, DCIS as purple and INV as gray.

Table 6.3: Comparison of the 4-class slide-level classification results when the training and the test sets involved the hand-crafted features [6] and the deep feature representations in two different experiments using the baseline MIMLL and the MimlHCRF algorithms. The evaluation criteria are: Hamming loss (HL), Jaccard index (JI), precision (P), recall (R) and f-measure (FM). The best result for each setting for the two sets of feature representations is marked in bold.

		HL	JI	P	R	FM
HAND-CRAFTED FEATURES	MimlNN	**0.254**	0.556	**0.661**	0.720	0.652
	MimlSvm	0.271	0.565	0.644	0.701	0.646
	MimlSvmMi	0.288	0.421	0.531	0.477	0.475
	M^3Miml	0.369	0.051	0.068	0.051	0.057
	MimlHCRF	0.288	**0.564**	0.655	**0.763**	**0.664**
DEEP FEATURES	MimlNN	0.220	0.630	0.695	0.737	0.692
	MimlSvm	0.258	0.540	0.703	0.588	0.610
	MimlSvmMi	0.271	0.438	0.610	0.438	0.494
	M^3Miml	0.263	0.379	0.542	0.379	0.432
	MimlHCRF	**0.191**	**0.665**	**0.815**	0.788	**0.754**
	$\text{MimlHCRF}_{1,2,3}$	0.249	0.627	0.665	0.860	0.715
	$\text{MimlHCRF}_{1,3}$	0.378	0.516	0.539	**0.931**	0.649
	$\text{MimlHCRF}_{1,2}$	0.448	0.352	0.464	0.541	0.450

The quantitative ROI-level classification results were obtained by comparing the most severe ROI-level prediction detected by the model to the most severe reference label, i.e. the consensus label, of the slide. Please note that the consensus ROIs were selected by experienced pathologists as example regions that corresponded to the most severe diagnosis in the slide. Hence, we compared the most severe diagnostic ROI-level label that the model predicted to the consensus label of the slide. Also note that the pathologists examined more regions when the slide had only Benign to make sure that no other region corresponded to a more severe diagnostic label. This caused more ROIs in slides which only had Benign, compared to the slides involving other classes, as pointed out in Table 6.2. This kind of imbalance in the number of ROIs could explain the tendency of the model making more Benign predictions in the absence of sufficient in-slide information.

The confusion matrix of the multi-class classification results is provided in Table 6.4 and the several statistics computed over the classification results were given in Table 6.5. The model was able to correctly classify the ROIs most of the

Table 6.4: The confusion matrix of the ROI-level classification results. The predictions were the most severe ROI-label in the slide and the targets were the consensus label of the same slide.

		Predicted			
		Benign	ADH	DCIS	INV
True	Benign	12	0	5	1
	ADH	11	6	1	0
	DCIS	4	1	12	1
	INV	1	0	2	2

Table 6.5: Class-specific statistics on the performance of ROI-level classification. The number of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) are given. Precision, recall (also known as true positive rate and sensitivity), false positive rate (FPR), specificity (also known as true negative rate) and F-measure are also shown.

Class	TP	FP	FN	TN	Precision	Recall/ Sensitivity	FPR	Specificity	F-measure
Benign	12	16	6	25	0.429	0.667	0.390	0.610	0.522
ADH	6	1	12	40	0.857	0.333	0.024	0.976	0.480
DCIS	12	8	6	33	0.600	0.667	0.195	0.805	0.632
INV	2	2	3	52	0.500	0.400	0.037	0.963	0.444

time. The highest precision value of 0.857 was observed on ADH predictions. The recall value for the cases with ADH suffered from the model's tendency towards Benign. On the other hand, cases with Benign were classified correctly with a high recall value of 0.667 but the model's being overly confident on Benign predictions led to a rather small precision value of 0.429. The performance of the model on DCIS was great in general with a precision value of 0.600 and a recall value of 0.667. Also, the model was able to correctly classify cases with INV and it mostly confused the cases with INV as DCIS due to frequently analyzing them together in the pathology forms of the slides. Finally, the overall accuracy of the MimlHCRF model on the 4-class ROI-level classification task was 0.542.

6.5 Discussion

This chapter introduced a multi-instance multi-label learning algorithm based on the flexible and powerful Hidden Conditional Random Field model for the multi-class classification of whole slide images and inference of dignostic labels

at ROI-level. The model was capable of simultaneous slide-level and ROI-level predictions through potential functions that were designed to model the associations between ROIs considering their spatial distribution within the slide, as well as the coherence and the correlations of each predicted ROI diagnosis and the multi-label slide-level diagnoses of the entire slide. The ROIs were sampled from the viewing behaviors of the pathologists and the slide-level labels were collected from the pathology forms they filled out. The correspondence between the ROIs and the slide-level labels were not known during the training to reflect the real-world clinical procedures. The slide-level labels included a more general set of 4 diagnoses from the original set of 14 classes of the diagnostic labels. These classes were specifically selected to reflect the clinical significance of the field, e.g. the differentiation of cases between ADH and DCIS, while also preserving both ends of the spectrum by including the cases from Benign and INV.

The feature representations for the variable-sized ROIs included both the traditional hand-crafted features such as color, texture and nuclear architecture and the features from a deep feature generator network involving convolutional architectures. Through several experiments, we evaluated the settings in which we compared the baseline MIMLL models and the proposed MimlHCRF model that were trained on the bags of deep feature representations through a deep generator network, to the same set of classifiers which were trained on the bags of hand-crafted traditional feature representations. In four of the five competing MIMLL models, experiments involving deep feature representations showed consistent improvement in all evaluation criteria against their counterparts with traditional feature representations. Overall, the best performing models for the two settings were the proposed MimlHCRF models.

MimlHCRF had the best performance when the feature representations of the ROIs came from the deep generator network, achieving the best classification result on each evaluation criterion, including the lowest Hamming loss; 0.191 and the highest precision and recall values; 0.815 and 0.788, respectively. The ROI-level performance of the same classifier had 0.542 accuracy on the 4-class classification task when the most severe ROI label in a slide was compared to the consensus label which corresponded to the label of the consensus ROIs of the

slide. The proposed model outperformed the existing MIMLL models on multi-class classification of whole slide images. Unlike its competition, our model was also capable of performing simultaneous ROI-level predictions without involving any kind of additional processes such as sliding windows-based methods.

Our study explored the complex dynamics of whole slide breast histopathology image analysis by combining several aspects of the field that were considered by the pathologists during their interpretations of slides in a real-world clinical setting such as taking the tissue structure of an ROI into account, incorporating the diagnostic cues in several ROIs into the slide-level decision, examining the spatial distribution of the ROIs within the cores present in the slide and considering the co-occurrence of different diagnostic labels. We investigated multi-class classification of whole slide breast histopathology images in a multi-instance multi-label learning scenario in which only the slide-level labels were present during the training and the feature representations of ROIs involved a feature generator network. As a future work, our study can benefit from attention networks for more accurate ROI discovery and can also be formulated as end-to-end stacked networks in a weakly supervised setting that can simultaneously generate deep feature representations while performing slide-level as well as ROI-level label predictions.

Chapter 7

Conclusion

Whole slide histopathology image analysis was traditionally performed on manually selected regions of interest. The process of selecting manual patches was not only costly but also did not reflect the routine clinical pathology practice that the pathologists followed during their interpretation of slides. With the advancements in the state-of-the-art convolutional neural networks, the research in histopathology image analysis has gradually embraced the capabilities of the field. However, the application of these methods in whole slide images has revolved around extracting small fixed-sized patches and training networks on those patches to try to uncover information about the entire slide, disregarding the fact that the slide-level diagnoses incorporated the diagnoses from variable number of variable-sized diagnostically relevant regions in the slide. In addition, analysis of histopathology images has involved benign or malignant cases corresponding to the cancer vs. non-cancer binary classification on the fixed-sized small patches, capturing limited clinical significance. To address these shortcomings, we introduced weakly supervised learning scenarios reflecting the real-world diagnostic process of the pathologists and a deep generator network that was able to learn feature representations at ROI-level, regardless of the shape or the size of the ROI. In this regard, this thesis had mainly three contributions for whole slide breast histopathology image analysis using machine learning techniques.

The first contribution included the formulation of the whole slide image classification problem considering the steps that the pathologists followed during the interpretation of the slides. This study reflected the real-world clinical setting that involved diagnoses from the pathology forms provided by the pathologists at slide-level where the diagnoses had clinical significance involving one or more of the several diagnostic labels. The selection of potentially diagnostically relevant regions inside a slide included the viewing records of the pathologists while they were interpreting the slide. We used certain actions to identify the ROIs in the slide which we referred as the instances of the bag, i.e. the slide, and the diagnoses in the pathology form of the slide provided by the pathologists were referred as the slide-level labels in the multi-instance multi-label learning scenario.

We formulated the weakly supervised learning framework which proposed the solution to the multi-instance multi-label learning problem by either transforming the classification problem into a multi-instance single-label learning task by assuming independence of labels or transforming the classification problem into single-instance multi-label learning task by embedding the bags into a new vector space. The instances were the traditional color, texture and nuclear features. This study marked the first work in the field of breast histopathology image analysis that involved multi-instance multi-label learning in a weakly supervised multi-class classification setting. Our experiments demonstrated that the weakly supervised classifiers could generalize well despite being only exposed to slide-level information during training.

The second main contribution of this thesis was the introduction of a novel deep feature representation method for ROIs in breast histopathology images. We introduced a feature generator network that learned deep convolutional feature representations for variable-sized ROIs in whole slide images. The significance of learning ROI-level feature representations stemmed from the fact that the slide-level diagnoses in a pathology form filled out by a pathologist corresponded to the diagnostic behaviour she observed on some of the diagnostically relevant regions in the whole slide image. Therefore, the discovery of such regions and learning their feature representations could lead to better identification of ROI-level labels as well as to better performance in slide-level classification. We designed a

feature generator network that involved the discovery of areas that were potentially more informative for the diagnosis in an ROI, and aggregated the feature representations of those areas using the properties of a deep convolutional network to finally form the deep feature representation of the ROI. The fine-tuning of the network was performed prior to the feature extraction process on sets of fixed-sized patches from the informative areas, each set extracted from an ROI and associated with the corresponding ROI-level label, from a separate fold in the data set involving no overlap with the training and test sets. The experiments demonstrated that the proposed feature representations could outperform the previous best efforts in ROI classification and the proposed approach was competitive at slide-level classification performance.

Following the baseline model, we investigated a more sophisticated multi-instance multi-label learning model that incorporated the complex relationships and associations in the data involving ROIs, ROI-level label predictions and slide-level diagnostic labels in a weakly supervised setting in the third contribution of this thesis. In this study, we incorporated a Hidden Conditional Random Field into the formulation of a multi-instance multi-label learning framework to perform simultaneous slide-level multi-class classification and ROI-level diagnostic label inference of whole slide breast histopathology images. The proposed model was capable of learning associations between ROI-level predictions based on their spatial distribution and it was also capable of learning from their individual structural information. Additionally, the model imposed coherence between slide-level and ROI-level label predictions, and it incorporated correlations in slide-level diagnoses from the pathology forms to improve the slide-level label predictions.

Our experiments included the deep feature generator network and the handcrafted color, texture and nuclear architecture features for the representations of the ROIs in separate experimental settings. We compared the performances of the baseline models and the proposed model when they were trained on bags involving both sets of feature representations, separately. The experimental results showed that the deep feature representations had more representational power in breast histopathology image classification within the multi-instance multi-label

learning framework. More importantly, we showed that our model could outperform the previous best efforts in multi-class classification of whole slide breast histopathology images in a weakly supervised setting. The proposed model could also infer ROI-level labels simultaneously while performing slide-level classification which the competition was not capable of.

Our extensive evaluations of various experimental settings confirmed that the combination of the representational power of convolutional networks for the ROIs with the powerful Hidden Conditional Random Field based multi-instance multi-label learning framework achieved the best results. The research in this thesis can be further extended as a future work in the form of an end-to-end framework that is capable of learning feature representations for the ROIs through a convolutional network while simultaneously performing slide-level multi-class classification as well as ROI-level label predictions from only slide-level information during training in a weakly supervised setting.

Bibliography

[1] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE Reviews in Biomedical Eng.*, vol. 2, pp. 147–171, October 2009.

[2] R. K. Jain, R. Mehta, R. Dimitrov, L. G. Larsson, P. M. Musto, K. B. Hodges, T. M. Ulbright, E. M. Hattab, N. Agaram, M. T. Idrees, and S. Badve, "Atypical ductal hyperplasia: interobserver and intraobserver variability," *Modern Pathology*, vol. 24, pp. 917–923, 2011.

[3] K. H. Allison, M. H. Rendi, S. Peacock, T. Morgan, J. G. Elmore, and D. L. Weaver, "Histological features associated with diagnostic agreement in atypical ductal hyperplasia of the breast: Illustrative cases from the B-Path study," *Histopathology*, vol. 69, pp. 1028–1046, 2016.

[4] J. G. Elmore, G. M. Longton, P. A. Carney, B. M. Geller, T. Onega, A. N. A. Tosteson, H. D. Nelson, M. S. Pepe, K. H. Allison, S. J. Schnitt, F. P. O'Malley, and D. L. Weaver, "Diagnostic concordance among pathologists interpreting breast biopsy specimens," *Journal of American Medical Association*, vol. 313, no. 11, pp. 1122–1132, 2015.

[5] C. Mercan, S. Aksoy, E. Mercan, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, "Multi-instance multi-label learning for whole slide breast histopathology," in *SPIE Medical Imaging Symposium, Digital Pathology Conference*, (San Diego, California), February 27–March 3, 2016.

[6] C. Mercan, S. Aksoy, E. Mercan, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, "Multi-instance multi-label learning for multi-class classification of

whole slide breast histopathology images," *IEEE Transactions on Medical Imaging*, vol. 37, no. 1, pp. 316–325, 2018.

[7] C. Mercan, S. Aksoy, E. Mercan, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, "From patch-level to roi-level deep feature representations for breast histopathology classification," in *SPIE Medical Imaging Symposium, Digital Pathology Conference*, (San Diego, California), February 17–21 2019.

[8] M. M. Dundar, S. Badve, G. Bilgin, V. Raykar, R. Jain, O. Sertel, and M. N. Gurcan, "Computerized classification of intraductal breast lesions using histopathological images," *IEEE Trans. on Biomedical Eng.*, vol. 58, pp. 1977–1984, July 2011.

[9] Y. Xu, J.-Y. Zhu, E. Chang, and Z. Tu, "Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 964–971, 2012.

[10] M. Kandemir and F. A. Hamprecht, "Computer-aided diagnosis from weak supervision: A benchmarking study," *Computerized Medical Imaging and Graphics*, vol. 42, pp. 44–50, 2015.

[11] M. Kandemir, C. Zhang, and F. A. Hamprecht, "Empowering multiple instance histopathology cancer diagnosis by cell graphs," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 228–235, 2014.

[12] E. Cosatto, P.-F. Laquerre, C. Malon, H.-P. Graf, A. Saito, T. Kiyuna, A. Marugame, and K. Kamijo, "Automated gastric cancer diagnosis on H&E-stained sections; training a classifier on a large scale with multiple instance machine learning," in *SPIE Medical Imaging*, vol. 867605, 2013.

[13] Y. Xu, J.-Y. Zhu, E. I.-C. Chang, M. Lai, and Z. Tu, "Weakly supervised histopathology cancer image segmentation and classification," *Medical Image Analysis*, vol. 18, no. 3, pp. 591–604, 2014.

[14] Y. Xu, L. Jiao, S. Wang, J. Wei, Y. Fan, M. Lai, and E. I.-C. Chang, "Multi-label classification for colon cancer using histopathological images," *Microscopy Research and Technique*, vol. 76, no. 12, pp. 1266–1277, 2013.

[15] Z. Han, B. Wei, Y. Zheng, Y. Yin, K. Li, and S. Li, "Breast cancer multi-classification from histopathological images with structured deep learning model," *Scientific Reports*, vol. 7, no. 1, p. 4172, 2017.

[16] F. G. Zanjani, S. Zinger, and P. H. N. de With, "Cancer detection in histopathology whole-slide images using conditional random fields on deep embedded spaces," in *SPIE Medical Imaging Symposium, Digital Pathology Conference*, p. 105810I, 2018.

[17] Y. Li and W. Ping, "Cancer metastasis detection with neural conditional random field," *arXiv preprint arXiv:1806.07064*, 2018.

[18] B. E. Bejnordi, G. Zuidhof, M. Balkenhol, M. Hermsen, P. Bult, B. van Ginneken, N. Karssemeijer, G. Litjens, and J. van der Laak, "Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images," *Journal of Medical Imaging*, vol. 4, no. 4, p. 044504, 2017.

[19] S. Mehta, E. Mercan, J. Bartlett, D. Weaver, J. Elmore, and L. Shapiro, "Learning to segment breast biopsy whole slide images," in *IEEE Winter Conference on Applications of Computer Vision*, pp. 663–672, 2018.

[20] S. Mehta, E. Mercan, J. Bartlett, D. Weaver, J. G. Elmore, and L. Shapiro, "Y-net: Joint segmentation and classification for diagnosis of breast biopsy images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2018.

[21] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "Breast cancer histopathological image classification using convolutional neural networks," in *International Joint Conference on Neural Networks*, pp. 2560–2567, 2016.

[22] N. Bayramoglu, J. Kannala, and J. Heikkilä, "Deep learning for magnification independent breast cancer histopathology image classification," in *International Conference on Pattern Recognition*, pp. 2440–2445, 2016.

[23] K. Das, S. P. K. Karri, A. G. Roy, J. Chatterjee, and D. Sheet, "Classifying histopathology whole-slides using fusion of decisions from deep convolutional network on a collection of random multi-views at multi-magnification," in *International Symposium on Biomedical Imaging*, pp. 1024–1027, 2017.

[24] 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.

[25] P. Kainz, M. Pfeiffer, and M. Urschler, "Segmentation and classification of colon glands with deep convolutional neural networks and total variation regularization," in *PeerJ*, p. e3874, 2017.

[26] 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 *Medical Imaging: Digital Pathology*, vol. 9041, p. 904103, 2014.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

[28] T. Araújo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polónia, and C. Aurélio, "Classification of breast cancer histology images using convolutional neural networks," *PloS one*, vol. 12, no. 6, p. e0177544, 2017.

[29] Y. S. Vang, Z. Chen, and X. Xie, "Deep learning framework for multi-class breast cancer histology image classification," in *International Conference Image Analysis and Recognition*, pp. 914–922, 2018.

[30] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2424–2433, 2016.

[31] T. H. Vu, H. S. Mousavi, V. Monga, G. Rao, and U. K. A. Rao, "Histopathological image classification using discriminative feature-oriented dictionary learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 3, pp. 738–751, 2016.

[32] J. Xu, L. Xiang, Q. Liu, H. Gilmore, J. Wu, J. Tang, and A. Madabhushi, "Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 1, pp. 119–130, 2016.

[33] A. Ng, "Sparse autoencoder," 2011.

[34] Y. Zheng, Z. Jiang, F. Xie, H. Zhang, Y. Ma, H. Shi, and Y. Zhao, "Feature extraction from histopathological images based on nucleus-guided convolutional neural network for breast lesion classification," *Pattern Recognition*, vol. 71, pp. 14–25, 2017.

[35] S. Manivannan, W. Li, J. Zhang, E. Trucco, and S. J. McKenna, "Structure prediction for gland segmentation with hand-crafted and deep convolutional features," *IEEE Transactions on Medical Imaging*, vol. 37, no. 1, pp. 210–221, 2018.

[36] N. V. Oster, P. A. Carney, K. H. Allison, D. L. Weaver, L. M. Reisch, G. Longton, T. Onega, M. Pepe, B. M. Geller, H. D. Nelson, T. R. Ross, A. N. A. Tosteson, and J. G. Elmore, "Development of a diagnostic test set to assess agreement in breast pathology: practical application of the Guidelines for Reporting Reliability and Agreement Studies (GRRAS)," *BMC Women's Health*, vol. 13, no. 3, pp. 1–8, 2013.

[37] J. G. Elmore, G. M. Longton, M. S. Pepe, P. A. Carney, H. D. Nelson, K. H. Allison, B. M. Geller, T. Onega, A. N. A. Tosteson, E. Mercan, L. G. Shapiro, T. T. Brunye, T. R. Morgan, and D. L. Weaver, "A randomized study comparing digital imaging to traditional glass slide microscopy for breast biopsy and cancer diagnosis," *Journal of Pathology Informatics*, vol. 8, no. 1, pp. 1–12, 2017.

[38] E. Mercan, S. Aksoy, L. G. Shapiro, D. L. Weaver, T. T. Brunye, and J. G. Elmore, "Localization of diagnostically relevant regions of interest in whole slide images," in *International Conference on Pattern Recognition*, pp. 1179–1184, 2014.

[39] E. Mercan, S. Aksoy, L. G. Shapiro, D. L. Weaver, T. T. Brunye, and J. G. Elmore, "Localization of diagnostically relevant regions of interest in whole slide images: A comparative study," *Journal of Digital Imaging*, vol. 29, pp. 496–506, August 2016.

[40] S. Doyle, M. Feldman, J. Tomaszewski, and A. Madabhushi, "A boosted bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies," *IEEE Trans. on Biomedical Eng.*, vol. 59, pp. 1205–1218, May 2012.

[41] A. Basavanhally, S. Ganesan, M. Feldman, N. Shih, C. Mies, J. Tomaszewski, and A. Madabhushi, "Multi-field-of-view framework for distinguishing tumor grade in ER+ breast cancer from entire histopathology slides," *IEEE Trans. on Biomedical Eng.*, vol. 60, pp. 2089–2099, August 2013.

[42] A. Ruifrok and D. Johnston, "Quantification of histochemical staining by color deconvolution," *Analytical and Quantitative Cytology and Histology*, vol. 23, no. 4, pp. 291–299, 2001.

[43] H. Xu, C. Lu, and M. Mandal, "An efficient technique for nuclei segmentation based on ellipse descriptor analysis and improved seed detection algorithm," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, pp. 1729–1741, September 2014.

[44] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in Neural Information Processing Systems*, pp. 561–568, 2002.

[45] L. Kaufman and P. J. Rousseeuw, "Clustering by means of medoids," in *Statistical Data Analysis Based on the L1-Norm and Related Methods* (Y. Dodge, ed.), pp. 405–416, North-Holland, 1987.

[46] G. Edgar, *Measure, Topology, and Fractal Geometry.* Springer Science & Business Media, 2007.

[47] Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li, "Multi-instance multi-label learning," *Artificial Intelligence*, vol. 176, no. 1, pp. 2291–2320, 2012.

[48] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognition*, vol. 37, no. 9, pp. 1757–1771, 2004.

[49] M.-L. Zhang and Z.-H. Zhou, "Multi-label learning by instance differentiation," in *AAAI Conference on Artificial Intelligence*, vol. 7, pp. 669–674, 2007.

[50] M.-L. Zhang and Z.-H. Zhou, "M3MIML: A maximum margin method for multi-instance multi-label learning," in *IEEE International Conference on Data Mining*, pp. 688–697, 2008.

[51] T. T. Brunye, P. A. Carney, K. H. Allison, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, "Eye movements as an index of pathologist visual expertise: A pilot study," *PLoS ONE*, vol. 9, no. 8, 2014.

[52] B. Gecer, "Detection and classification of breast cancer in whole slide histopathology images using deep convolutional networks," Master's thesis, Bilkent University, 2019.

[53] B. E. Bejnordi, M. Balkenhol, G. Litjens, R. Holland, P. Bult, N. Karssemeijer, and J. A. W. M. van der Laak, "Automated detection of DCIS in whole-slide h&e stained breast histopathology images," *IEEE Trans. on Medical Imaging*, vol. 35, pp. 2141–2150, September 2016.

[54] M. L. McHugh, "Interrater reliability: the Kappa statistic," *Biochemia Medica: Biochemia Medica*, vol. 22, no. 3, pp. 276–282, 2012.

[55] X. Zhou, K. Yu, T. Zhang, and T. S. Huang, "Image classification using super-vector coding of local image descriptors," in *European Conference on Computer Vision*, pp. 141–154, 2010.

[56] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "From generic to specific deep representations for visual recognition," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 36–45, 2015.

[57] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 447–456, 2015.

[58] A. Bansal, X. Chen, B. Russell, A. Gupta, and D. Ramanan, "Pixelnet: Representation of the pixels, by the pixels, and for the pixels," *arXiv preprint arXiv:1702.06506*, 2017.

[59] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computing Research Repository*, vol. abs/1409.1556, 2015.

[60] H. Abdi and L. J. Williams, "Principal component analysis," *Wires Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[61] M. M. Kokar and J. A. Tomasik, "Data vs. decision fusion in the category theory framework," in *International Conference on Information Fusion*, pp. 3–15, 2001.

[62] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention*, pp. 234–241, 2015.

[63] E. Mercan, *Digital Pathology: Diagnostic Errors, Viewing Behavior and Image Characteristics*. PhD thesis, University of Washington, 2017.

[64] M. M. Dundar, S. Badve, V. C. Raykar, R. K. Jain, O. Sertel, and M. N. Gurcan, "A multiple instance learning approach toward optimal classification of pathology slides," in *International Conference on Pattern Recognition*, pp. 2732–2735, 2010.

[65] S. Kumar and M. Hebert, "Discriminative random fields: A discriminative framework for contextual interaction in classification," in *IEEE International Conference on Computer Vision*, pp. 1150–1157, 2003.

[66] C. Sminchisescu, A. Kanaujia, and D. Metaxas, "Conditional models for contextual human motion recognition," *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 210–220, 2006.

[67] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1848–1852, 2007.

[68] Z.-J. Zha, X.-S. Hua, T. Mei, J. Wang, G.-J. Qi, and Z. Wang, "Joint multi-label multi-instance learning for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[69] L. Maaten, M. Welling, and L. Saul, "Hidden-unit conditional random fields," in *International Conference on Artificial Intelligence and Statistics*, pp. 479–488, 2011.

[70] S. B. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell, "Hidden conditional random fields for gesture recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1521–1527, 2006.

[71] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt, "Hidden conditional random fields for phone classification," in *European Conference on Speech Communication and Technology*, pp. 1117–1120, 2005.

[72] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1717–1724, 2014.

[73] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[74] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

[75] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[76] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. V. D. Laak, B. V. Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[77] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[78] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721–741, 1984.

[79] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[80] T. Tieleman, "Training restricted Boltzmann machines using approximations to the likelihood gradient," in *International Conference on Machine learning*, pp. 1064–1071, 2008.

[81] X. He, R. S. Zemel, and M. Á. Carreira-Perpiñán, "Multiscale conditional random fields for image labeling," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. II–II, 2004.

[82] N. Dong, M. Kampffmeyer, X. Liang, Z. Wang, W. Dai, and E. Xing, "Reinforced auto-zoom net: Towards accurate and fast breast cancer segmentation in whole-slide images," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 317–325, Springer, 2018.

[83] B. Gecer, S. Aksoy, E. Mercan, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, "Detection and classification of cancer in whole slide breast histopathology images using deep convolutional networks," *Pattern Recognition*, vol. 84, pp. 345–356, 2018.

[84] S. Godbole and S. Sarawagi, "Discriminative methods for multi-labeled classification," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 22–30, 2004.

[85] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, 2014.