DISCRIMINATIVE REPRESENTATIONS FOR HETEROGENEOUS IMAGES AND MULTIMODAL DATA

Heather D. Couture

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Computer Science.

Chapel Hill
2019

Approved by:

Marc Niethammer

Alex Berg

J.S. Marron

Charles M. Perou

Stephen Pizer

ABSTRACT

Heather D. Couture: Discriminative Representations for Heterogeneous Images and Multimodal Data
(Under the direction of Marc Niethammer)

Histology images of tumor tissue are an important diagnostic and prognostic tool for pathologists. Recently developed molecular methods group tumors into subtypes to further guide treatment decisions, but they are not routinely performed on all patients. A lower cost and repeatable method to predict tumor subtypes from histology could bring benefits to more cancer patients. Further, combining imaging and genomic data types provides a more complete view of the tumor and may improve prognostication and treatment decisions. While molecular and genomic methods capture the state of a small sample of tumor, histological image analysis provides a spatial view and can identify multiple subtypes in a single tumor. This intra-tumor heterogeneity has yet to be fully understood and its quantification may lead to future insights into tumor progression.

In this work, I develop methods to learn appropriate features directly from images using dictionary learning or deep learning. I use multiple instance learning to account for intra-tumor variations in subtype during training, improving subtype predictions and providing insights into tumor heterogeneity. I also integrate image and genomic features to learn a projection to a shared space that is also discriminative. This method can be used for cross-modal classification or to improve predictions from images by also learning from genomic data during training, even if only image data is available at test time.

ACKNOWLEDGMENTS

First and foremost, I want to express my gratitude to my advisor, Marc Niethammer. He has provided immense guidance, encouragement, and support throughout the last seven years, starting even before I applied to the graduate program at UNC. He made the decision to return to school for a Ph.D. an easy one. He gave me the freedom to explore different research directions and helped refine my writing skills. He challenged me to think about theory and design clear experiments. He accommodated me in working remotely, allowing for work/life balance while raising a family. I appreciate his dedication to both teaching and advising.

I also want to thank the other professors on my committee. Steve Marron provided a continuous source of insightful advice and feedback on my research. I appreciate his focus on model interpretability and eagerly await future contributions from his group in this area. Chuck Perou has been essential in contributing knowledge of breast cancer genomics and guiding this project in a clinically important direction. Steve Pizer provided my first introduction to medical image analysis and has continued to contribute valuable and insightful feedback on my research presentations and this dissertation. Alex Berg provided valuable feedback on my research through his expertise in computer vision and machine learning.

I am also grateful to a number of other collaborators. Melissa Troester provided extensive knowledge in breast cancer research, guidance towards clinically meaningful directions, and access to and support with the Carolina Breast Cancer Study data. Lindsay Williams co-authored a journal paper with me, contributing essential statistical analysis and epidemiological insights; she also provided guidance on the breast cancer data set. Joseph Geradts and David Eberhard contributed their pathology expertise and insights. Susan Wei got me started in working with histology data and collaborated on melanoma research along with Nancy Thomas and Jayson Miedema.

Funding for this research was provided by a Royster fellowship, a grant from the Lineberger Comprehensive Cancer Center, and the Center for Computer Integrated Systems for Microscopy

and Manipulation (funded by NIH). A Tesla K40 was donated by Nvidia.

Finally, I am thankful for the support of my family. My husband Rob has provided continuous love, support, and encouragement. He has kept me grounded and challenges me to be better every day. Daniela and Lucas keep life exciting - their laughs, smiles, and hugs make my day. I am grateful to my parents for doing everything that parents do and much more. Thank you for raising me in an environment in which I knew that any career direction I chose was possible.

TABLE OF CONTENTS

LIST OF TABLES ix

LIST OF FIGURES x

CHAPTER 1: Introduction 1

1.1 Computer Science Motivations 3

1.2 Motivations in Cancer Research 4

1.3 Thesis Statement and Contributions 6

1.4 Overview of Chapters 8

CHAPTER 2: Representing Histology Images 9

2.1 Overview of Representations 10

2.2 Stain Normalization 13

2.3 Dictionary Learning 13

2.4 Deep Transfer Learning 18

2.5 Experiments 19

2.5.1 Data Sets 20

2.5.2 Implementation Details 20

2.5.3 Unsupervised Feature Comparison 21

2.5.4 Stain Normalization with Dictionary or Deep Transfer Learning 23

2.5.5 Hierarchical Task-driven Dictionary Learning 24

2.6 Discussion 28

CHAPTER 3: Multiple Instance Learning for Heterogeneous Images with an SVM 29

3.1 Related Work 30

3.2 Aggregation Functions for Single Instance Learning 34

3.3 Generating Instances 34

3.4 Iterative Multiple Instance Learning 35

3.5 Quantile Aggregation 36

3.6 Sample Weighting by Class 38
3.7 Experiments 39
3.7.1 Data Sets 39
3.7.2 Implementation Details 40
3.7.3 Classification Results 41
3.7.4 Sample Weighting by Class 43
3.7.5 Statistical Validation 44
3.7.6 Visualization of Heterogeneity 44
3.8 Discussion 44
CHAPTER 4: Multiple Instance Learning for Heterogeneous Images with a CNN 47
4.1 Background 48
4.2 Multiple Instance Learning with a CNN 49
4.3 Multiple Instance Aggregation 50
4.4 Training with Multiple Instance Augmentation 51
4.5 Experiments 52
4.5.1 Data Set 52
4.5.2 Implementation Details 52
4.5.3 MI Augmentation and the Importance of MI Learning 53
4.5.4 MI Aggregation 54
4.5.5 CNN Architecture 55
4.5.6 Pre-trained vs. Fine-tuned CNN 56
4.5.7 Heterogeneity 57
4.6 Discussion 58
CHAPTER 5: Integrating Image and Genomic Features with Task-Driven Deep CCA 60
5.1 Introduction 60
5.2 Background: CCA and Deep CCA 62
5.3 Task-driven Deep CCA 64
5.4 Related Work 73

5.5 Experiments 75
5.5.1 Implementation Details 76
5.5.2 Synthetic Examples with MNIST Split 76
5.5.3 Cross-modal Classification on Real Data 81
5.5.4 CCA for Regularization on CBCS 85
5.6 Discussion 86
CHAPTER 6: Conclusions 88
6.1 Summary of Contributions 88
6.2 Software and Data Availability 96
6.3 Future Work 97
6.3.1 Training a CNN for Histopathology 97
6.3.2 Multimodal Deep Learning 98
6.3.3 Deep Learning Data Challenges 99
6.3.4 Cancer Research 100
6.4 Closing Remarks 102
APPENDIX A: Experimental Validation of Breast Tumor Histology Classification 103
A.1 Introduction 103
A.2 Methods 103
A.3 Results 105
A.4 Discussion 111
REFERENCES 116

LIST OF TABLES

2.1 Patient-level AUC for different unsupervised feature representations and classifiers ...22

2.2 RGB histology images vs. stain normalization with dictionary learning ...23

2.3 RGB histology images vs. stain normalization with deep transfer learning...24

2.4 Patch-level classification accuracy ...25

2.5 Patient-level classification accuracy...26

3.1 Classification accuracy using features from AlexNet and VGG16...43

3.2 ER status and Basal vs. non-Basal with and without grade weighting...44

4.1 Average classification accuracy for different types of MI aggregation ...55

4.2 Classification accuracy for different CNN architectures ...55

4.3 Pre-trained vs. fine-tuned CNN ...56

5.1 Cross-modal classification results on CBCS ...82

5.2 Cross-modal classification results on TCGA-BRCA ...85

5.3 Classification accuracy for predicting from images only at test time. ...86

A.1 Patient and tumor characteristics for the image analysis training and test set ...106

A.2 Grading agreement between pathologists and image analysis...108

A.3 Impact of weighting by grade on accuracy, sensitivity, and specificity of ER status...109

A.4 Classification performance for intrinsic subtype, ROR-PT, and histologic subtype ...110

A.5 Patient and tumor characteristics associated with inaccuracy of ER status ...112

LIST OF FIGURES

1.1 Example of a tissue microarray with a single core magnified 2

2.1 Stain normalization. 14
2.2 Overview of hierarchical dictionary learning 14
2.3 Overview of Zero-phase Component Analysis 15
2.4 Dictionary elements learned on nuclei-centered and dense patches. 22
2.5 Relevance maps for a sample image 27

3.1 Intra-tumor and intra-subtype heterogeneity 30
3.2 Related work for MI 31
3.3 Overview of iterative MI method 37
3.4 Process for classifying images with multiple instances and quantile aggregation 38
3.5 Recognition rate for different MI methods on the BreaKHis data set 41
3.6 AUC and classification accuracy for different MI methods on the CBCS data set 42
3.7 ROC plots for MIL-median with VGG16 43
3.8 Predicted tumor heterogeneity across four H&E cores from a single patient 45

4.1 MI augmentation 48
4.2 MI framework with a CNN 50
4.3 Classification accuracy for different cropped image sizes 54
4.4 t-SNE plots of pre-trained vs. fine-tuned CNN features. 57
4.5 Visualization of instance predictions 58
4.6 Predicted heterogeneity for grade and genomic subtype 58

5.1 Deep CCA network architectures 65
5.2 Sum correlation vs. classification accuracy for DCCA and SoftCCA. 77
5.3 Batch size vs. classification accuracy on MNIST split 78
5.4 Training set size and input dimension vs. classification accuracy on MNIST split 78
5.5 t-SNE plots for CCA methods on MNIST split. 80
5.6 t-SNE plots for CCA methods on CBCS 84

A.1 Histogram of predicted grade vs. pathologist-classified 107
A.2 Bee swarm plot of predicted grade vs. pathologist-classified 108

CHAPTER 1: INTRODUCTION

Cancer is a heterogeneous disease, comprising multiple subtypes associated with distinctive morphology, genomics, clinical presentations, responses to treatment, and outcomes. Understanding its diverse nature is critical in tailoring treatments to each patient. Although there are distinct subtypes of cancer, significant intra-tumor heterogeneity still exists in many individual cancers, posing a challenge for diagnosis and treatment. To date, most research has focused on identifying and characterizing distinct subtypes [Paik et al., 2004; Parker et al., 2009], while only more recent interest has been directed at understanding the effects of heterogeneity [Alizadeh et al., 2015; McGranahan and Swanton, 2015].

Understanding cancer requires the use of a heterogeneous mix of data types: clinical, histology, radiology, genomics, and proteomics, among others. Each modality provides a complementary view of the same lesion. I target histologic and genomic analysis of breast tumor tissue. Histology images show microscopic views of tissue and enable pathologists to diagnose disease and assess prognosis. However, interpretation by pathologists is limited by its subjectivity, speed, and the capability of humans to discriminate complex properties. Through image analysis, I form automated methods for assessing the prognosis and subtype of tumor tissue, which could lead to better treatment decisions for cancer. Although the focus of this dissertation is on breast cancer, the techniques are generalizable and could provide powerful quantitative analysis methods for many other cancer types and diseases in general.

Tissue samples are obtained by biopsy, sectioned, and placed on a glass slide. Prior to imaging, they are typically stained to increase the contrast of structures of interest. Hematoxylin and eosin (H&E) are the most widely used stains for histological diagnosis; they turn nuclei blue and cytoplasm pink. At 20x magnification, even small structures such as individual nuclei are visible. The tissue samples I work with are from a tissue microarray (TMA), in which small representative core samples (typically four per tumor) are extracted and imaged on a single slide; Figure 1.1 shows an example. These core samples are about 2500 pixels in diameter,

whereas a full slide is often 50,000 pixels or more in width. Adjacent slices of tissue are stained to identify other tumor properties such as estrogen, progesterone, and HER2 receptor status. Other core samples are sent for genomic analysis, providing the expression level of a set of genes and the genomic subtype. The PAM50 subtype has been shown clinically-relevant and is associated with the tumor receptor status [Parker et al., 2009].

Tumor tissue is far from homogeneous. Different types of heterogeneity may be present: 1) Although tissue cores are taken from tumor tissue, they may contain adjacent stroma (connective) or adipose (fatty) tissue, each with a different appearance and each may or may not be connected with the discrimination task at hand. 2) A single genomic subtype is assigned to each sample, but the tumor may contain a mix of subtypes. 3) Tumors of different histologic types may belong to the same genomic subtype, resulting in a varied appearance for each subtype. Methods to target each of these challenges will be explored.

This dissertation studies ways to learn representations directly from histology images, capture intra-tumor heterogeneity, and integrate image and genomic data to form more predictive models for subtyping, grading, and prognosis. Feature learning is an integral component to each chapter. It is used to represent image patches and sub-regions of an image as instances in multiple instance (MI) learning. A dictionary of image patches can provide a basis for representing a set of heterogeneous components and aids interpretation as only a small number of dictionary elements are used to reconstruct a sample. More powerful, but less interpretable, features are obtained with a Convolutional Neural Network (CNN). Discriminative models are also essential to each of the frameworks that I work with. Some methods are faster to train,

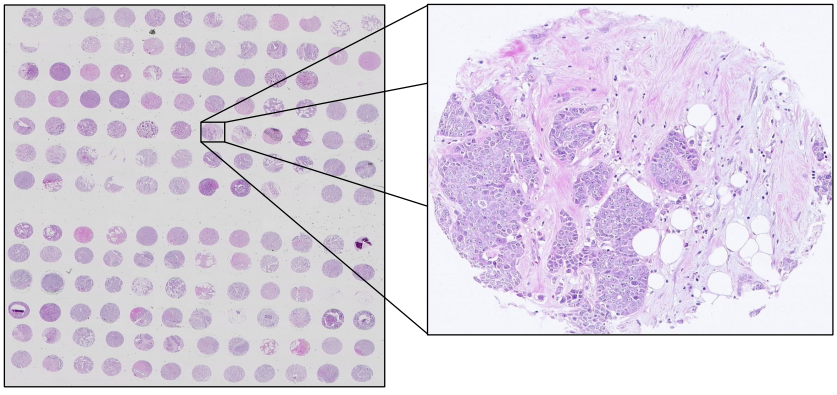

Figure 1.1: Example of a tissue microarray with a single core magnified.

some are top performers when given a lot of data, and others excel in the high-dimensional low sample size (HDLSS) setting. Classifiers with each of these characteristics will be employed and adapted to the unique properties of histologic images. When combined, feature learning with a discriminative model provides a powerful way to make predictions from heterogeneous data.

1.1 Computer Science Motivations

Discriminative Image Representations. Traditional approaches to image analysis involved hand-crafting domain-specific features to describe the color, shape, or texture of images [Julesz, 1981; Gotlieb and Kreyszig, 1990; Miedema et al., 2012]. However, hand-crafted features are difficult to develop and to transfer to new applications. Somewhat more generic hand-engineered features were later invented to characterize local regions of an image [Lowe, 2004; Bay et al., 2008; Dalal and Triggs, 2005], but are not adapted to the needs of a specific image type or application. More recent work in dictionary learning and deep learning forms features directly from the images [Mairal et al., 2009; Le et al., 2012b]. However, most recent advances have focused on forming discriminative features from small images using deep learning [Krizhevsky et al., 2012; He et al., 2016; Szegedy et al., 2017]. Extensions of these methods are necessary to handle large, heterogeneous images and to find subtle differences between image classes that are not visually distinguishable to domain experts.

Multiple Instance Learning on Heterogeneous Images. Breaking large images into small regions for making class predictions is the first step in accommodating heterogeneous images [Chen and Wang, 2004; Li and Vasconcelos, 2015]. While prior work has explored different ways to aggregate predictions into a single image classification for some problem types [Andrews et al., 2002; Song et al., 2013; Li and Vasconcelos, 2015; Carbonneau et al., 2017], more general aggregation approaches are needed for heterogeneous images. Further, instance aggregation techniques in a deep learning setup can help CNNs adapt to this weakly labeled image scenario.

Multimodal Data Analysis. A variety of cross-modal analysis methods exist, mostly based on Canonical Correlation Analysis (CCA), enabling the projection of two modalities of data to a shared space [Hotelling, 1936; Andrew et al., 2013]. This shared space can then be used for cross-modal classification or other data analysis tasks. While a great deal of effort has focused on maximizing the correlation between modalities in this shared space [Wang et al., 2015a; Chandar et al., 2016; Chang et al., 2018], there has been very little focus on ensuring that the shared space remains discriminative. Task-driven deep CCA methods are needed to find a projection that is both highly correlated and discriminative. By ensuring robustness in the HDLSS setting, such a method can be applicable to medical applications and can further data interpretation efforts.

1.2 Motivations in Cancer Research

Diagnosis and Prognosis. Pathologists examine biopsied or resected tissue to identify the presence of a tumor and to characterize multiple features in order to assess tumor aggressiveness. Accurate diagnosis and assessment of tumors is fundamental in providing appropriate and timely treatment. By improving the accuracy of diagnosis, we can reduce over-treatment of benign lesions and under-treatment of malignant ones. Better survival prediction can help to determine how closely to monitor patients and which patients should be offered entrance into clinical trials.

Visual Assessment by Pathologist. Tissue samples are typically graded by a pathologist to assess prognosis. For breast cancer, a more favorable outcome is expected if there is a higher percentage of tubule and gland formation, low nuclear pleomorphism (the irregularity of nuclear size and shape), and a low mitotic count [Tavassoli and Devilee, 2003]. Although they are well trained in examining tissue, making such assessments can be highly variable among pathologists, particularly for intermediate grade tumors [Longacre et al., 2006; Salles et al., 2008; Broekaert et al., 2010]. Automated computational methods can make these assessments more repeatable and also capture more complex properties than a pathologist can assess visually. By learning features that pathologists may not have thought to assess or that may be too complex to characterize visually, we can provide new insights into factors driving tumor progression.

Genomic Subtypes. Although visual examination of tissue by a pathologist is typically used for diagnosis and grading, genomic subtyping can further guide treatment decisions for the goal of personalized medicine. For breast cancer, intrinsic subtypes are determined through the PAM50 molecular genomic assay, dividing tumors into five classes [Parker et al., 2009]. These subtypes have different prognoses and respond to treatment differently [Parker et al., 2009].

Heterogeneity. Genomic subtypes are assigned from a single sample of tissue; they cannot provide a spatial view of the tumor. Although some tumors are mostly homogeneous, others are very heterogeneous, likely due to their branched evolution [Hiley and Swanton, 2014]. This heterogeneity can lessen the predictive ability when multiple types are present in a single tumor, as the patient would fall into a different subtype dependent on which part of the tumor was sampled. The implications on prognosis and targeted therapies are not yet well understood [Alizadeh et al., 2015]. Histology gives us the ability to examine the spatial heterogeneity of the tumor in a way that gene expression cannot.

Computational Methods for Subtypes. Recent advances in prognostication have relied on molecular and genomic methods that are costly and not routinely performed on all patients who could benefit. Predicting genomic properties from H&E histology alone could identify patients who are most like to benefit from further genomic testing. Further, genomic data is not always available, such as in low resource settings or due to insufficient tissue. In these situations, computational image-based methods could provide a lower cost substitute.

Multimodal Models. Histologic image features and genomics provide two complementary views of tumors. By combining the two, a more complete picture of tumor prognosis and treatment models can be developed. Each alone has been shown to inform certain decisions made by doctors, but an integrated model can provide a more predictive analysis. Initial efforts for breast and prostate cancer show the potential of multimodal methods for outcome prediction [Yuan et al., 2012; Lee et al., 2015].

Interpretation. Through my analysis and in learning features directly from the data, I also provide an interpretation mechanism. Visualizing features and locating regions of tumor that

most contributed to a prediction can create a teaching mechanism for pathologists.

Other Medical Applications. Although my focus is on H&E histology data sets, the techniques described in this work are not specific to this type of image. They could provide powerful quantitative analysis methods for many other cancer types and staining protocols and also many other diseases exhibiting heterogeneity, such as chronic obstructive pulmonary disease (COPD) [Mannino, 2002; Agusti et al., 2010] and Alzheimer's disease [Lambert and Amouyel, 2007].

1.3 Thesis Statement and Contributions

Learned representations for histology images of tissue can capture both intra- and inter-tumor heterogeneity, enabling discriminative models for tumor properties. Combining these image features with data from other modalities such as genomics in a task-driven model can provide insight into the shared tumor properties and further improve predictions. These computational techniques using discriminative features can provide a lower cost and more repeatable alternative to molecular methods and insight into tumor heterogeneity.

The contributions of this dissertation in Computer Science include:

1) Discriminative representations for histology images using dictionary learning or deep transfer learning. The dictionary learning method is task-driven to discover subtle differences between classes and hierarchical to capture architectural properties. The deep transfer learning method validates the use of pre-trained CNN features for discriminative tasks on non-RGB images.

2) Multiple instance learning methods for handling large, heterogeneous images with an SVM on any type of feature set or with a CNN for end-to-end training. An iterative SVM-based method learns the latent instance labels given a particular assumption on instance label aggregation. Alternatively, a more general MI method uses the quantile function for pooling and learns how to aggregate instance predictions. This quantile method works with either an SVM or end-to-end training with a CNN. An MI augmentation technique is used while training the CNN and enables the exploration of single instance and MI learning on a continuous spectrum. Insight into both SVM- and CNN-based methods is

provided by visualizing the predictions of each instance.

3) A set of multimodal methods to find a shared space that is also discriminative. This set of deep CCA models can be used for cross-modal classification and in gaining insight into the shared components of two modalities. They bring the CCA projection into the network itself in different ways, enabling end-to-end training to optimize both the correlation between modalities and the task-driven goal.

4) Techniques for deep learning on problems traditionally viewed as "small data." Solutions in this regime include deep transfer learning, multiple instance learning on large images, multi-task learning, and appropriate regularization.

Further contributions to the application area of breast cancer research include:

A) Methods to capture biologically-relevant features by operating on the H&E stain intensities extracted from histology images. These methods do not rely on hand-crafted features and are shown to produce more accurate predictions. The feature learning methods are also easily transferable to other cancer types.

B) A low cost and repeatable method for predicting histopathological, molecular, and genomic properties of tumors from H&E histology. Experimental validation shows that the classification accuracies achieved are comparable to the inter-rater agreement of pathologists and of alternative ways of assessing tumor properties. Further, these methods showed success on predicting molecular and genomic properties from H&E histology - something not previously known to be possible from H&E alone.

C) A mechanism to find predicted tumor heterogeneity from H&E histology. While genomic subtyping methods assess a very small region of tumor, imaging provides a spatial view. Predicting tumor subtype from smaller regions across the image provides a view of heterogeneity. Future work will be necessary to validate the predicted tumor heterogeneity and assess its association with patient outcome.

1.4 Overview of Chapters

This dissertation addresses aspects of heterogeneity in tumor tissue and of data types in order to form more predictive models for tumor properties. In Chapter 2, I study representation methods for histology images by comparing hand-crafted features, dictionary learning, and deep learning. Dictionary learning is further extended into a task-driven framework, and pre-trained CNN features are applied to histology data. Next, I focus on intra-tumor heterogeneity. Although genomic subtypes and other biomarkers are typically assessed at the tumor level, many tumors contain a mix of subtypes. Multiple instance learning enables the formation of more predictive models, predicts which parts of a tumor are associated with each subtype, and provides insight into subtype heterogeneity. Chapter 3 focuses on SVM-based methods that can use any feature set, while Chapter 4 brings MI learning into a CNN for end-to-end training. In Chapter 5, I explore the integration of heterogeneous sources of data to form a discriminative representation. I propose a set of four methods to create a task-driven deep CCA model that projects both modalities into a shared space that is also discriminative. Finally, Chapter 6 wraps up with a summary of contributions and a discussion of future work.

CHAPTER 2: REPRESENTING HISTOLOGY IMAGES

Appropriately representing images is the first and most critical step in applying automated analysis. The features must capture the important properties of the image for the chosen discrimination task. In the case of predicting diagnosis, prognosis, or subtype, the class differences can be very subtle. Machine learning methods perform such tasks by first representing images by a vector of features. Traditionally, these were hand-crafted features describing the color, shape, or texture of the image [Julesz, 1981; Gotlieb and Kreyszig, 1990; Tuceryan and Jain, 1998; Miedema et al., 2012] or hand-engineered features capturing properties of local image patches [Lowe, 2004; Bay et al., 2008; Dalal and Triggs, 2005]. A classifier then maps each feature vector to a class prediction, such as normal versus disease [Lepistö et al., 2003; Varma and Zisserman, 2005; Csurka et al., 2004].

Histological image analysis presents many challenges due to variations in staining and biological heterogeneities [Niethammer et al., 2010]. Each tissue type has specialized structures [Young et al., 2013], making hand-crafted features developed for one type difficult to apply to another. Tumors from different genomic subtypes may also appear similar, requiring features that capture their subtle differences. Rather than engineering specific features for each data type and task, a representation can be learned directly from the data [Varma and Zisserman, 2007; Coates et al., 2010; Coates and Ng, 2011]. Dictionary learning and deep learning are two such methods that I explore in this chapter.

This chapter will first outline the related work in this area (Section 2.1). I then present hierarchical task-driven dictionary learning in Section 2.3. Section 2.4 discusses transfer learning with a pre-trained CNN. Finally, Section 2.5 validates each method on histology images of tumors.

2.1 Overview of Representations

Hand-crafted Cell Features. Several previous studies have utilized automated processing of H&E stained breast tumors to identify image features associated with tissue types or outcome. They generally follow an approach of first segmenting nuclei, then characterizing color, texture, shape, and spatial arrangement properties of cells and nuclei [Miedema et al., 2012; Cooper et al., 2012; Chang et al., 2011]. A simple averaging of cell and nuclear properties in a region of tissue has limitations due to its focus on local properties of cells. Further, these hand-crafted features are time-consuming to develop and do not adapt easily to new data sets. Prior work on automated grading addresses mitotic count [Veta et al., 2015], nuclear atypia [Khan et al., 2015], and tubule formation [Basavanhally et al., 2011] individually; however, the latter two require a nuclear segmentation that is also difficult to adapt to new data sets.

Hand-engineered Patch Descriptors. Other more general feature descriptors developed for other types of images have also been applied to histology. Cruz-Roa et al. compare greyscale image patches, the Scale Invariant Feature Transform (SIFT), the Discrete Cosine Transform (DCT), and Local Binary Patterns (LBP) [Cruz-Roa et al., 2011, 2014]. Although these and other features have great utility on many image analysis problems, they are not optimal for histology or many other fine-grained image classification tasks [Cruz-Roa et al., 2013, 2014].

Dictionary Learning. The current trend in image analysis is towards raw image patches as the input to feature learning. Learning a dictionary for the sparse reconstruction of image patches has produced successful results on histology [Han et al., 2011; Zhou et al., 2014]. This sparse coding method has also been shown to produce superior image classification results in comparison to other encoding methods when used in a single-level dictionary learning framework [Coates and Ng, 2011]. Additional modifications to improve the discrimination capability of the dictionary involve learning a separate dictionary for each class [Vu et al., 2015] or combining the reconstruction and classification error into a single objective function [Ranzato and Szummer, 2008; Jiang et al., 2013; Mairal et al., 2012]. The former has been successful for classes that are easier to distinguish, while the latter better captures fine-grained differences between classes. Mairal et al. applied task-driven dictionary learning to improve recognition of hand-written

digits [Mairal et al., 2012]. I extend it to a hierarchical dictionary learning framework for classifying large images.

Dictionary learning is computationally-intensive in both learning the dictionary during training and computing coefficients at test time because it requires an optimization process rather than a feed-forward computation. For this reason, I implemented the encoding function for processing on a Graphics Processing Unit (GPU). The complexity of the computation is also a problem during training; therefore, I implemented the hierarchical framework in a greedy manner, learning one layer at at time.

Deep Learning. Deep learning is a method of learning a hierarchy of features where the higher level concepts are built on the lower level ones [LeCun et al., 2015]. Automatically learning these abstract features enables the system to learn complex functions mapping an input to an output without the need for hand-crafted features. In comparison to dictionary learning, deep learning is much more efficient and can thus be scaled to larger models. Highly optimized software is commonly available for training and testing convolutional neural networks (CNNs) on a GPU. While the encoding function for each layer is simpler than for dictionary learning, the power comes from stacking many layers into a larger model with end-to-end training.

The largest downside of deep learning is the limited interpretability of these large models. Methods have been developed to identify which regions of an image are important for classification [Simonyan et al., 2013; Bach et al., 2015] or visualize individual features [Zeiler and Fergus, 2014], but this is still very much an active research area. In Chapters 3 and 4 I will develop heatmaps to identify which regions of an image are most associated with its class. Gaining insight into individual features is not tackled in this dissertation but will be important in the future, particularly as deep learning is increasingly applied to medical applications.

Deep learning has been applied to histology as various forms of neural networks [Le et al., 2012a; Nayak et al., 2013; Chang et al., 2013] and CNNs [Cruz-Roa et al., 2013]. Recent successes in deep learning have been shown in recognizing hand-written digits and objects [Le et al., 2012b; Krizhevsky et al., 2012; Sermanet et al., 2014; Szegedy et al., 2015; Simonyan and Zisserman, 2015]. Simpler deep learning architectures have also been applied in classifying large histology images [Le et al., 2012a; Nayak et al., 2013; Chang et al., 2013; Cruz-Roa et al.,

2013] and for specific tasks such as mitosis detection [Cireşan et al., 2013; Veta et al., 2015], tissue segmentation [Xu et al., 2016], and segmentation and detection of a number of tissue structures [Janowczyk and Anant, 2016].

Many state-of-the-art CNN models are trained with tens or hundreds of millions of labeled images. In the medical domain, expert annotations are expensive and patient samples are scarce. Training a large model on a small data set may result in overfitting - the model performs well with the data it was trained on but gives poor results on newly presented data [Cawley and Talbot, 2010]. This is because the model has too many parameters relative to the amount of labeled data. Predicting diagnosis, subtype, prognosis, or other such complex classes uses patient-level labels, making it much more difficult to obtain large quantities of labeled data. End-to-end training of a CNN will be addressed in Chapter 4. In this chapter, I use a pre-trained CNN for transfer learning.

Transferred Deep Features. To accommodate the limitations of small data sets, deep learning models trained on more general image data sets can be transferred to specific applications. Typically, the network is transferred at some intermediate layer, and either the network is fine-tuned on the new data or the transferred layer of features is used to train a new classifier. The former has shown a lot of success for fine-grained classification tasks that have larger amounts of labeled data [Yosinski et al., 2014; Azizpour et al., 2014; Zhang et al., 2015]. The latter has been applied to tasks with smaller data sets and can still outperform hand-engineered features [Zhang et al., 2015; Codella et al., 2015; Donahue et al., 2014]. Transferability is less for distant tasks, particularly for higher layers due to their specialization [Yosinski et al., 2014], making this technique even more difficult to apply to a specialized image set such as histology.

Lu et al. compare techniques with training using randomly initialized weights for classification tasks on computed tomography [Lu et al., 2016]. For larger networks especially, transferring and fine-tuning was found to be important. Cruz-Roa et al. have applied transfer learning without fine-tuning to histology in comparing a deep CNN trained on ImageNet with a shallower CNN trained on a different histology data set than the one tested on [Cruz-Roa et al., 2015]. The smaller network trained on similar data was the top performer. Shouno et al. have shown success in fine-tuning the upper two layers of a CNN on a data set with only 12,000 training

examples [Shouno et al., 2015]. Each of these applications still has many more labeled samples than the data sets that I work with. I study the use of a large pre-trained CNN in extracting features for classifying histology images.

Applications on Histology. Significant advances in image analysis for histology have shown promise for tumor detection [Cruz-Roa et al., 2013], metastatic cancer detection in lymph nodes [Wang et al., 2016a], mitosis detection [Veta et al., 2015; Cireşan et al., 2013], tissue segmentation [Xu et al., 2016], and segmentation and detection of a number of tissue structures [Janowczyk and Anant, 2016]. However, all of the previous successes of deep learning from H&E images have focused on detecting image-based properties that pathologists routinely assess visually. Using deep learning to predict complex properties that are not visually apparent to pathologists, such as receptor status or genomic subtype, has not been previously described. I tackle predicting both types of tumor characteristics.

2.2 Stain Normalization

Staining of tissue samples is commonly used to highlight structures of interest, with hematoxylin and eosin as the most commonly-used set for histological diagnosis. Hematoxylin turns nuclei blue and eosin turns cytoplasm pink. Standardization of slide appearance can help to counter variations due to slide fading, differing stain colors, and the variety of microscopes and imaging equipment used. Color and intensity normalization helps to minimize these variations by estimating the stain vectors for hematoxylin and eosin and normalizing each image.

I use the method by Niethammer et al. [Niethammer et al., 2010] that uses prior information on the absorption coefficients for each stain in order to provide support for images with sparsely distributed nuclei. The image is decomposed into the components of each individual stain using color deconvolution. The resulting stain intensity channels are then used as input to the rest of my algorithm. An example is shown in Figure 2.1.

2.3 Dictionary Learning[1]

[1]The methods presented in this section were presented at the IEEE International Symposium on Biomedical Imaging in 2015 [Couture et al., 2015].

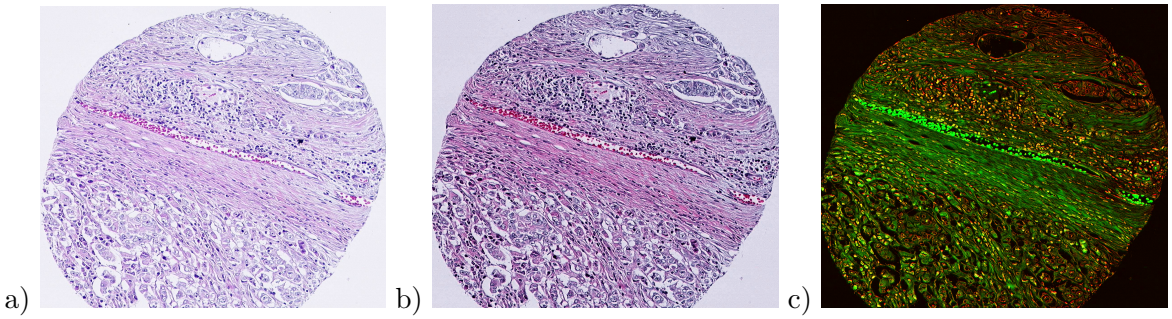

Figure 2.1: Stain normalization: a) original H&E image, b) stain normalized, c) stain intensities with hematoxylin in the red channel, eosin in the green channel, and the residual in the blue channel.

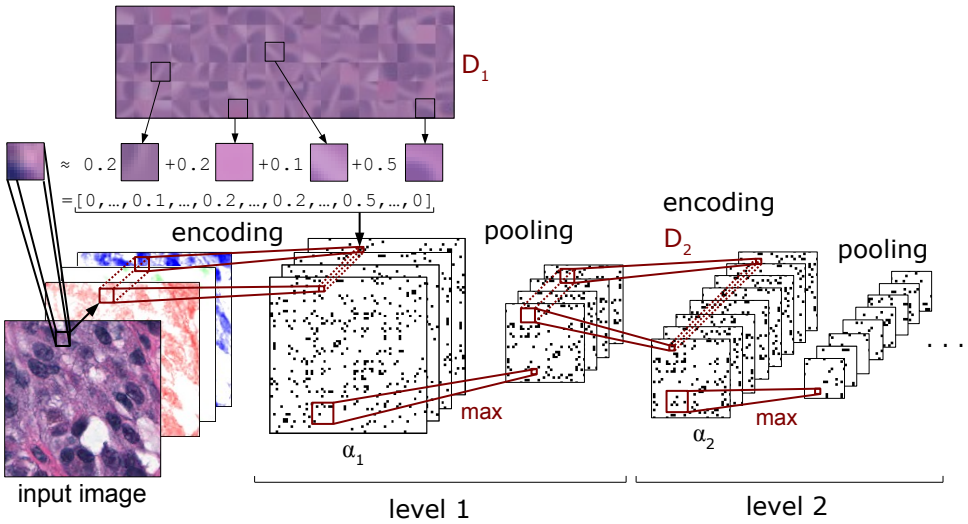

Figure 2.2: Overview of hierarchical dictionary learning: Images are first color normalized and the hematoxylin, eosin, and residual stain channels extracted. Each image patch is encoded using a dictionary. Following encoding, a max pooling operation downsizes the image. By alternating encoding and pooling layers, a hierarchy of features is formed.

Dictionary learning based on sparse coding can learn representations for histology images. A dictionary is learned from image patches in the training set, which is then used to encode patches in novel images. This section outlines the steps to learn hierarchical task-driven dictionaries and apply them to encode images for classification. Whitening is first used to decorrelate image patches. Figure 2.2 provides an overview of this hierarchical process of image encoding.

Whitening. Dictionary learning operates on square patches extracted from training images. Adjacent pixels are highly correlated. I first apply mean centering and a Zero-phase Component Analysis (ZCA) whitening step to reduce the redundancy of individual patches by making the

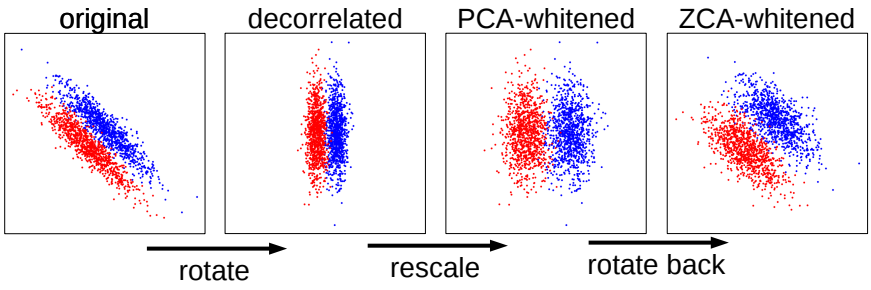

Figure 2.3: Overview of Zero-phase Component Analysis (ZCA). Computed using SVD, three operations are performed: a rotation to decorrelate the data, a rescaling of each axis, and a rotation back to the original space.

features uncorrelated and to give each feature a similar variance [Hyvärinen and Oja, 2000; Mairal et al., 2014]. Whitening has previously been found to improve image classification accuracy when applied as preprocessing [Coates et al., 2010]. This centering and whitening process is applied prior to encoding for each level of the hierarchy.

A set of n image patches, each flattened into a d-dimensional vector, is stored in matrix $X \epsilon \mathbb{R}^{d \times n}$ ($n >> d$) and mean centered. Whitening applies a transformation $\tilde{X} = UX$ to make $\tilde{X} \epsilon \mathbb{R}^{d \times n}$ orthonormal: $\tilde{X}\tilde{X}^T = I$. The covariance matrix of X is computed as $C = \frac{1}{n-1} XX^T$. Any matrix $U \epsilon \mathbb{R}^{d \times d}$ that satisfies the condition $U^T U = C^{-1}$ whitens the data; however, U is only defined up to a rotation, so it is not unique. I first apply PCA whitening to decorrelate the features, followed by a rotation back to the original space. PCA whitening uses the eigendecomposition of covariance matrix C: $U_{PCA} = \Lambda^{-1/2} V^T$ for $\Lambda = \text{diag}(\sigma_1, ..., \sigma_d)$ and $V = [v_1, ..., v_1]$, where (σ_i^2, v_i) are the eigenvalue, eigenvector pairs of C. ZCA uses the transformation $U_{ZCA} = V \Lambda^{-1/2} V^T$, in which PCA whitening is first applied, followed by a rotation back to the original space. While PCA is commonly used to reduce the data dimensionality, ZCA typically keeps all d dimensions. Adding the rotation V brings the whitened data $\tilde{X}$ as close as possible to the original input data X [Kessy et al., 2015]. An overview of ZCA whitening is shown in Figure 2.3.

Unsupervised Dictionary Learning. I use sparse coding to learn a dictionary of features to represent image patches. The elastic net formulation looks for a small number of dictionary elements that, through a linear combination, can reconstruct a given image patch. I selected the elastic net rather than the Lasso model because the coefficients are better behaved in the case

where dictionary elements are similar. Given whitened input data $\{x_1, ..., x_n\}, x_i \epsilon \mathbb{R}^d$, the goal is to compute a dictionary $D \epsilon \mathbb{R}^{d \times k}$ and coefficients $\{\alpha_1, ..., \alpha_n\}$ such that the reconstruction error $\sum_{i=1}^{n} ||x_i - D\alpha_i||^2$ is minimized and the coefficients α are sparse.

If the dictionary D is known, the coefficients α can be computed by optimizing

$$\alpha^*(x, D) = \underset{\alpha}{\operatorname{argmin}} \frac{1}{2}\|x - D\alpha\|_2^2 + \lambda_1\|\alpha\|_1 + \lambda_2\|\alpha\|_2^2. \tag{2.1}$$

The ℓ_1 norm encourages sparsity in α, and the ℓ_2 norm adds stability in the case of correlated variables. Due to the computationally intensive nature of evaluating the elastic net, I perform this on a GPU.

Initially, the dictionary D must be learned from the data. It is computed from a set of whitened image patches $\{x_1, ..., x_n\}$ after first initializing with random patches. I use a similar elastic net formulation in which both the coefficients α and the dictionary D must be learned:

$$D, \alpha = \underset{D,\alpha}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2}\|x_i - D\alpha_i\|_2^2 + \lambda_1\|\alpha_i\|_1 + \lambda_2\|\alpha_i\|_2^2$$

such that for each column of D, $\|D_{:,j}\|_2^2 \leq 1$. Although this cost function is not convex, it can be split into the sub-problems of optimizing for the dictionary D given the coefficients α, and optimizing for the coefficients α given the dictionary D. Each of these sub-problems is convex, so the typical approach is to alternate the two steps until convergence. I use the online batch implementation by Mairal et al. [Mairal et al., 2009].

Task-Driven Dictionary Learning. The discriminating power of the dictionary can be improved by incorporating image label information into the dictionary learning framework [Mairal et al., 2012]. By minimizing the logistic loss, I learn a linear discriminant for two classes based on the sparse encodings of individual image patches. Although I focus on binary classification here, the logistic could be replaced with the softmax function to generalize to multiple classes.

I initialize the dictionary using the unsupervised dictionary learning procedure detailed in the previous section. An initial linear classifier is learned using logistic regression on the encodings $\alpha^*(x, D)$ of a set of training patches $\{x_1, ..., x_n\}$. The classifier is defined by a

separating hyperplane w such that if $w^T\alpha^*(x, D) + w_0 > 0$, patch x is predicted to belong to class 2, and class 1 otherwise. The logistic function $1/[1 + e^{-(w^T\alpha^*(x,D)+w_0)}]$ predicts a probability indicating how likely the patch is to belong to class 2.

In improving the dictionary and classifier, the logistic loss objective function I use is as follows:

$$f(D, w) = \min_{w,D} \sum_{i=1}^{n} \log[1 + e^{-y_i(w^T\alpha^*(x_i,D)+w_0)}] + \frac{\nu}{2}\|w\|_2^2$$

where y_i is the class label (-1 or 1) associated with each patch x_i, w defines the hyperplane separating the two classes, $\alpha^*(x, D)$ is defined in (2.1), and parameter ν controls the regularization. I optimize this objective by stochastic gradient descent, updating D and w as

$$D \leftarrow D - \gamma\nabla_D f(D, w) \qquad w \leftarrow w - \gamma\nabla_w f(D, w)$$

where γ is the learning rate, and $\nabla_w f(D, w)$ and $\nabla_D f(D, w)$ are calculated from the logistic loss function $f(D, w)$ using $\nabla_D\alpha^*(x, D)$ derived by Mairal et al. [Mairal et al., 2012].

Hierarchy of Features. Now that I can form dictionaries of learned features and use them to encode images, I turn to the problem of forming a feature hierarchy to capture more abstract and larger scale properties. After densely encoding every patch in an image, a max pooling operation is applied in which, for each $m \times m$ region, I take the maximum encoded value for each feature. This has the effect of providing local translation invariance and downsizing the representation to enable capture of larger-scale properties by the next level. Encoding and max pooling operations are alternated to form a feature hierarchy.

Classification. At this point, each image is represented by a set of sparse encodings of features from each level of the hierarchy and I must predict the image-level class. I can apply the logistic regression classifier to each image patch or summarize the encodings themselves and train a new classifier. I compare four image-level classification methods:

1) The mean of the patch probabilities over the image.

2) The sum of the log of patch probabilities (equivalent to multiplying probabilities).

3) A new logistic regression classifier to operate on quantile functions summarizing patch probabilities.

4) A Support Vector Machine (SVM) to operate on histograms of the patch encodings (equivalent to a mean pool of the encodings).

For the first two options, I found it to work best if a threshold to separate the two classes is learned on the training data. I experiment on each of these strategies in Section 2.5.5.

Summary. This section presented dictionary learning in a task-driven and hierarchical framework. The task-driven component increases the discriminability of dictionary elements while the hierarchical part captures larger-scale and more abstract features, such as tissue architecture. Task-driven dictionary learning has previously been applied to small images [Mairal et al., 2012]; I applied it to patches within a larger image and proposed four aggregation schemes for image classification. I also extended task-driven dictionary learning from a single layer to a multi-layer hierarchy.

2.4 Deep Transfer Learning

CNNs consist of convolution filters applied to small patches of the image, followed by a non-linearity and often a data reduction or pooling layer [LeCun et al., 2015]. Convolutional and pooling layers are typically alternated to extract features and provide local translational invariance, respectively [Krizhevsky et al., 2012; Simonyan and Zisserman, 2015]. Fully connected layers may be added on top of the convolutional network, and a softmax regression layer is applied for classification. Backpropagation is used to learn the network parameters. Similar to human visual processing, the low level filters detect small structures such as edges and blobs. Intermediate layers capture increasingly complex properties like shape and texture. The top layers of the network are able to represent object parts like faces or bicycle tires. The network weights are learned from data, creating discriminating features at multiple levels of abstraction. There is no need to hand-craft features.

Many of the top-performing CNN architectures such as AlexNet [Krizhevsky et al., 2012] and VGG16 [Simonyan and Zisserman, 2015] were trained on the ImageNet data set, which

consists of 1.2 million images from 1000 categories of objects and scenes. Although ImageNet contains a vastly different type of image, CNNs trained on this data set have been shown to transfer well to other data sets [Oquab et al., 2014; Razavian et al., 2014; Yosinski et al., 2014], including those from biomedical applications [Wang et al., 2016a; Tajbakhsh et al., 2016]. The lower layers of a CNN are fairly generic, while the upper layers are much more specialized. The lower layers only capture smaller-scale features, which do not provide enough discriminating ability, while the upper layers are so specific to ImageNet that they do not generalize well to histology. Intermediate layers are both generalizable and discriminative for other tasks.

In transferring to histology, I search for the layer that transfers best to a particular task. I extracted the output from each layer over each image at full resolution to form a set of features for the image. Taking the mean of each feature over the tissue region (excluding the background), then forms a single feature vector for each image - essentially a weighted global mean pool with weights of one where there is tissue and zero for the background. When multiple images are available for each patient, I further average across images.

The ImageNet data set on which the CNNs were pre-trained consists of RGB photographs of scenes and objects. H&E histology images contain a much more limited set of colors and appearances with their focus on the blue/purple color of nuclei with hematoxylin staining, the pink color of surrounding tissue with eosin staining, and the remainder mostly white. While the stain-normalized RGB images of histology provide one view, I also experiment with the decomposition into hematoxylin, eosin, and residual. I use these three channels as input by first subtracting a per-channel mean across the data set and then extracting features with a pre-trained CNN. Experimental results comparing feature learning strategies and assessing stain normalization for preprocessing are in Sections 2.5.3 and 2.5.4, respectively.

2.5 Experiments

This section details some experiments to validate the presented feature learning methods. I will first discuss the data sets used and implementation details for each method. Experiments will then be presented in the following order: 1) a comparison of unsupervised feature methods, 2) a comparison of dictionary learning and deep transfer learning with and without stain

normalization, and 3) hierarchical task-driven dictionary learning.

2.5.1 Data Sets

Melanoma. The melanoma data set consists of whole slide images in which a pathologist has annotated an average of eight regions containing tumor. 31 of these samples contain varying degrees of dysplastic nevi (benign), while 21 contain melanoma.

SPECS. My second data set contains breast tumor samples from a Washington University cohort of patients [Parker et al., 2009]. These take the form of a tissue microarray with two cores per patient; they were imaged at the University of British Columbia. I predict the subtype of the 43 Basal and 42 Luminal A samples.

CBCS. This data set consists of 1713 patient samples from the Carolina Breast Cancer Study, Phase 3 [Troester et al., 2018]. There are typically four cores per patient (5970 cores total), with each core image having a diameter of around 2400 pixels. I predict Basal vs. non-Basal intrinsic subtype, ER positive vs. negative, and grade 1 vs. 3.

2.5.2 Implementation Details

Deep Transfer Learning. Features were extracted using the AlexNet CNN pre-trained on ImageNet [Krizhevsky et al., 2012]. The mean output of the fourth convolutional layer over the tissue region was taken for each image and further averaged over all images for each patient.

Dictionary Learning. A patch size of 17×17 and dictionary size of 256 was used for unsupervised dictionary learning. I selected λ_1 from 0.25, 0.5, 1.0, and 2.0 as the value that produced the best patch classification accuracy through cross-validation on the training set. I set λ_2 to $\lambda_1/10$ to add some stability to the model, while keeping the ℓ_1 norm as the main mode of regularization. The mean of the computed coefficients from overlapping patches was taken for each image and further averaged over all images for each patient.

Hierarchical Task-driven Dictionary Learning. Patch sizes of 9×9, 5×5, and 3×3 and dictionary sizes of 128, 192, and 256 were used for the three levels, respectively, with a

3×3 max pool for each. Dictionary learning requires setting the regularization parameters λ_1 and λ_2 (Section 2.3). I selected λ_1 from 0.25, 0.5, 1.0, and 2.0 as the value that produced the best patch classification accuracy through cross-validation on the training set. I once again set λ_2 to $\lambda_1/10$ to add some stability to the model, while keeping the ℓ_1 norm as the main mode of regularization. The logistic loss of task-driven dictionary learning requires a regularization parameter ν (Section 2.3). I also learned this from the data as the value from 10^{-6} to 10^{1} that produced the greatest patch classification accuracy. During learning, patches were randomly selected from each image and were randomly flipped and/or rotated to add more variety to the data. A learning rate γ of 10^{-5} was found to work with my data sets in combination with a batch size of $500000/N$ patches from each image, where N is the number of training images; 60, 20, and 15 cycles through the training set were used for the three levels respectively.

Classifier Hyperparameters. Hyperparameters for logistic regression, SVM, and DWD were learned through cross-validation on the training set to maximize the area under the ROC curve (AUC).

2.5.3 Unsupervised Feature Comparison

Using the SPECS data set, I compared different methods of unsupervised feature learning. Feature representations for the training images were learned without the class labels, followed by encoding the images in the learned representation and training a classifier using the class labels. The AUC was computed as a measure of accuracy for each method and is presented in Table 2.1 as the mean over two rounds of five-fold cross-validation. The hand-crafted features used are described by Miedema et al. and capture the size, shape, stain intensity, texture, and local spatial arrangement of cells and nuclei [Miedema et al., 2012]. Dictionary learning was run on nuclei-centered patches and dense overlapping patches; Figure 2.4 shows the dictionary elements that were learned. A pre-trained CNN was also compared by taking the outputs from the fourth convolutional layer of AlexNet [Krizhevsky et al., 2012].

These results indicate that a linear SVM or logistic regression is most suitable as a classifier. Hand-crafted features were consistently outperformed by learned features. Dictionary learning on dense patches was more successful than on only nuclei-centered patches. Intermediate lay-

Method	Log. Reg.	SVM - Linear	SVM - RBF	DWD
Hand-crafted features	78.9% (3.2)	77.8% (2.7)	57.3% (4.0)	72.8% (2.2)
Nuclei-centered patches	81.2% (2.0)	79.4% (1.7)	66.1% (4.5)	75.5% (3.5)
Dense patches	84.5% (2.0)	85.5% (2.4)	63.1% (3.5)	79.9% (2.3)
AlexNet Conv4	83.2% (2.9)	82.5% (3.2)	71.6% (3.9)	81.1% (3.4)

Table 2.1: Patient-level AUC results for different unsupervised feature representations and classifiers on the SPECS data set, with the standard error in brackets.

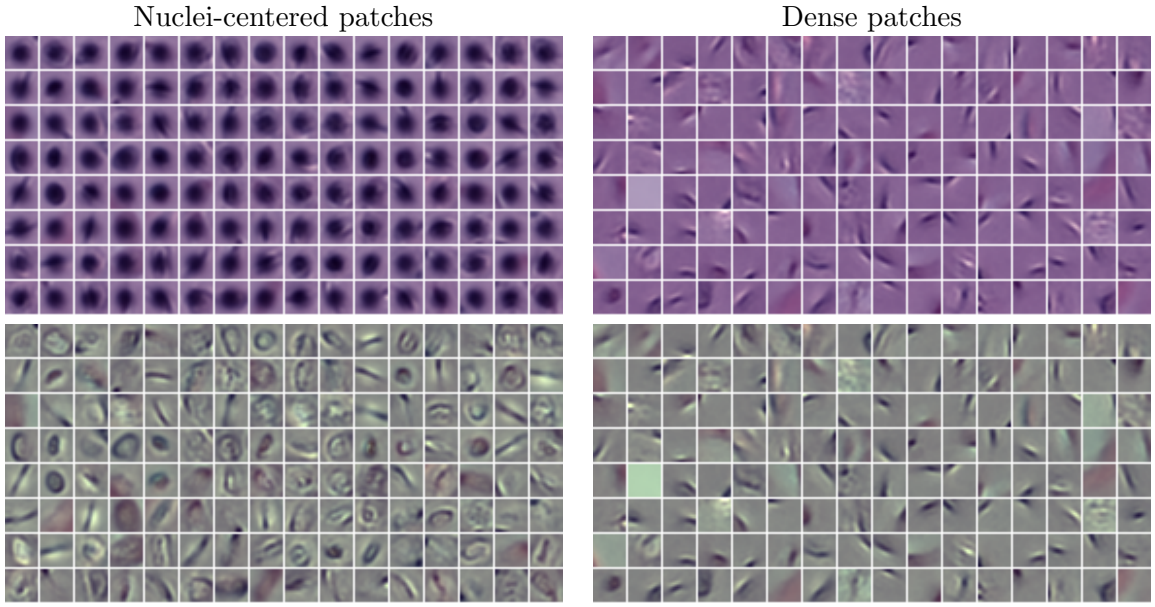

Figure 2.4: Subset of elements from unsupervised dictionary learning framework. Top row: dictionary elements with whitening and mean centering reversed. Bottom row: dictionary elements with whitening reversed, but without mean centering reverted.

Input image	AUC			Accuracy		
	Basal	ER	Grade	Basal	ER	Grade
Original RGB	81.0 (0.8)	84.3 (0.9)	90.5 (0.7)	79.0 (1.2)	79.7 (0.9)	82.8 (1.0)
Normalized RGB	81.7 (1.0)	85.0 (0.9)	91.1 (1.0)	78.0 (1.0)	79.0 (1.2)	82.1 (1.3)
Stain channels	82.2 (1.4)	86.0 (0.8)	92.7 (0.9)	79.5 (1.2)	80.5 (0.9)	84.8 (1.0)

Table 2.2: AUC and classification accuracy on CBCS using dictionary learning on the original unnormalized RGB images, stain normalized RGB images, and extracted stain channels (hematoxylin, eosin, and residual). Standard error is in brackets.

ers from a pre-trained CNN showed very promising results, although not quite as strong as dictionary learning on dense patches. This is in line with Cruz-Roa et al.'s conclusion that a simple model trained on the desired data type is better than a more complex one trained on a disparate data set [Cruz-Roa et al., 2015].

2.5.4 Stain Normalization with Dictionary or Deep Transfer Learning

I studied stain normalization as a preprocessing step to feature learning. The experiments in this section use the CBCS data set and 5-fold cross validation. The mean and standard error across the five folds is reported.

Table 2.2 compares the classification performance with dictionary learning on the original RGB histology images, stain normalized RGB, and the extracted stain channels (hematoxylin, eosin, and residual). Dictionary learning with the stain normalized RGB images was generally better than with the original histology images. Using the stain channels however, consistently outperformed the other two input image types.

Next, I studied stain normalization when used with deep transfer learning by taking the outputs from the fourth convolutional layer of AlexNet averaged over the tissue region [Krizhevsky et al., 2012]. I compare classification performance on the CBCS data set for the original RGB images, stain normalized RGB, and different permutations of extracted stain channels. The purpose of this experiment was to gain insight into whether the power of deep transfer learning is due to certain key features from the pre-trained CNN or simply the set of features as a whole. Table 2.3 shows that no permutation of stain channels performed measurably better than the others, or than the stain normalized RGB images. If the particular feature representation produced by AlexNet had certain features that were key in classifying histology, then a different

Input image	AUC			Accuracy		
	Basal	ER	Grade	Basal	ER	Grade
Original RGB	78.4 (1.7)	81.9 (0.9)	90.6 (1.4)	77.5 (1.3)	76.7 (0.6)	81.9 (1.3)
Normalized RGB	80.7 (1.3)	85.7 (1.1)	92.8 (0.5)	78.4 (1.4)	79.2 (1.0)	84.8 (0.6)
Stain channels 012	78.5 (1.5)	85.1 (0.8)	93.3 (0.9)	77.8 (1.5)	79.8 (0.9)	84.2 (1.5)
Stain channels 021	78.0 (1.4)	83.3 (0.7)	92.4 (0.9)	77.7 (1.3)	79.2 (0.6)	83.2 (1.1)
Stain channels 102	77.3 (1.7)	84.2 (1.2)	92.7 (1.2)	77.3 (1.2)	78.5 (1.0)	84.2 (1.0)
Stain channels 120	79.8 (1.4)	84.0 (0.8)	91.7 (1.1)	78.6 (1.5)	79.0 (0.9)	83.5 (0.9)
Stain channels 201	76.7 (1.2)	83.7 (1.1)	92.5 (0.6)	77.7 (1.2)	78.4 (0.9)	84.5 (0.9)
Stain channels 210	77.8 (1.8)	84.9 (0.2)	92.7 (0.9)	77.4 (1.4)	79.7 (0.8)	83.2 (0.4)

Table 2.3: AUC and classification accuracy with deep transfer learning for the original unnormalized RGB images, stain normalized RGB images, and different channel permutations on extracted stain channels (hematoxylin, eosin, and residual). Standard error is in brackets.

representation of the image (e.g., by permuting the color channels and again extracting features with AlexNet) would no longer exhibit the same discriminability. This is shown to not be the case, indicating that it is not the individual features captured by the pre-trained CNN providing the discriminative power but the space that they span.

The stain normalized RGB images and extracted stain channels do, however, outperform the original unnormalized RGB images for ER status and grade, but not for Basal vs. non-Basal. This might indicate that ER status and grade distinctions are based on color, but Basal vs. non-Basal is learned from other properties. The dictionary learning results in Table 2.2 also show a similar outcome in that the increase in classification performance due to stain normalization is greater for ER status and grade than for Basal vs. non-Basal.

Tables 2.2 and 2.3 also provide a comparison of unsupervised dictionary learning and deep transfer learning on the larger CBCS data set. While Basal vs. non-Basal was more successful with dictionary learning, ER status and grade 1 vs. 3 performed about equally well with the two methods. Task-driven dictionary learning (this chapter) and fine-tuning a CNN (Chapter 4) can provide further improvements for classification. A larger CNN such as VGG16 [Simonyan and Zisserman, 2015] can also provide more discriminative features.

2.5.5 Hierarchical Task-driven Dictionary Learning[2]

[2]The results presented in this section were presented at the IEEE International Symposium on Biomedical Imaging in 2015 [Couture et al., 2015].

	Melanoma vs. nevi		Breast subtype	
	U	TD	U	TD
Level 1	55.2%	**59.0%**	50.7%	**52.0%**
Level 2	59.8%	**63.9%**	56.4%	**58.0%**
Level 3	59.0%	**70.0%**	51.1%	**54.6%**

Table 2.4: Patch-level classification accuracy comparing unsupervised dictionaries (U) with task-driven dictionaries (TD) for a 3-level hierarchy.

I assessed both unsupervised and task-driven dictionary learning as a hierarchy by comparing the classification accuracy on the melanoma and SPECS data sets.

Classification Results. In order to assess the importance of both the task-driven and hierarchical components of my model, I set up experiments to measure the patch-level and patient-level classification accuracy using 5-fold cross-validation. Although prediction accuracy on patients is expected to be much greater than that on local patches, both provide a means of validation and the latter is important for model interpretation.

First, using the logistic regression classifiers trained during task-driven dictionary learning, I computed the patch-level classification accuracy before and after the task-driven learning process (Table 2.4). Both data sets showed a consistent improvement of task-driven dictionaries over unsupervised ones. The melanoma data set also showed a consistent improvement from level 1 to 3, with a small decrease in the unsupervised dictionary performance for level 3. The breast subtype results showed a significant drop in performance for level 3 for both methods. This data set is much more complex and poses a more challenging problem. Algorithm parameters such as patch size and dictionary size likely need to be better tuned to get better results on this data set.

I also measured the patient-level classification accuracy using each of the methods detailed in Section 2.3 (Table 2.5). This showed a fairly consistent improvement from level 1 to 3 for the first three methods that summarize the image using the patch classifier. However, the breast subtype results were not as consistent as those for melanoma, likely due to the reasons already mentioned for the patch-level results. The task-driven dictionary method outperformed the unsupervised dictionary on the melanoma data set, but only in some settings on the breast subtype data set. The SVM method on feature histograms performed well across the different

	Melanoma vs. nevi		Breast subtype	
	U	TD	U	TD
1. Mean of patch probabilities				
Level 1	65.5%	53.6%	61.5%	59.3%
Level 2	82.9%	84.4%	64.9%	64.6%
Level 3	84.5%	88.5%	70.1%	62.1%
2. Sum of log of patch probabilities				
Level 1	63.3%	74.7%	64.6%	64.2%
Level 2	84.7%	86.5%	62.4%	63.4%
Level 3	82.7%	88.4%	67.5%	58.6%
3. Logistic regression on quantile of patch probabilities				
Level 1	59.6%	67.5%	72.9%	66.4%
Level 2	79.1%	78.5%	65.7%	63.7%
Level 3	81.1%	82.4%	63.5%	65.6%
4. Linear SVM on histogram of features				
Level 1	86.5%	84.7%	69.8%	71.3%
Level 2	84.7%	84.5%	70.6%	65.4%
Level 3	82.9%	84.4%	68.3%	70.2%

Table 2.5: Patient-level classification accuracy comparing unsupervised dictionaries (U) with task-driven dictionaries (TD) for a 3-level hierarchy using the four different methods described in Section 2.3.

levels but did not show an improvement from higher levels.

Model Interpretation I now turn to the problem of identifying which regions of an image are most associated with each class. Using the logistic regression classifier trained on patches, I predicted the probability that an individual patch belonged to each class (Section 2.3). I formed a colormap in which blue indicates class 1 has a higher probability, red indicates class 2, and white is neutral. This is shown for a melanoma image in Fig. 2.5 and compares the results from unsupervised and task-driven dictionaries for a 3-level hierarchy. These results show that the task-driven dictionary produced a slightly higher confidence in classification for levels 1 and 2, as indicated by slightly more red coloring and less blue. The confidence in melanoma also increased up the levels; however, level 3 showed a decrease in confidence for the unsupervised dictionary.

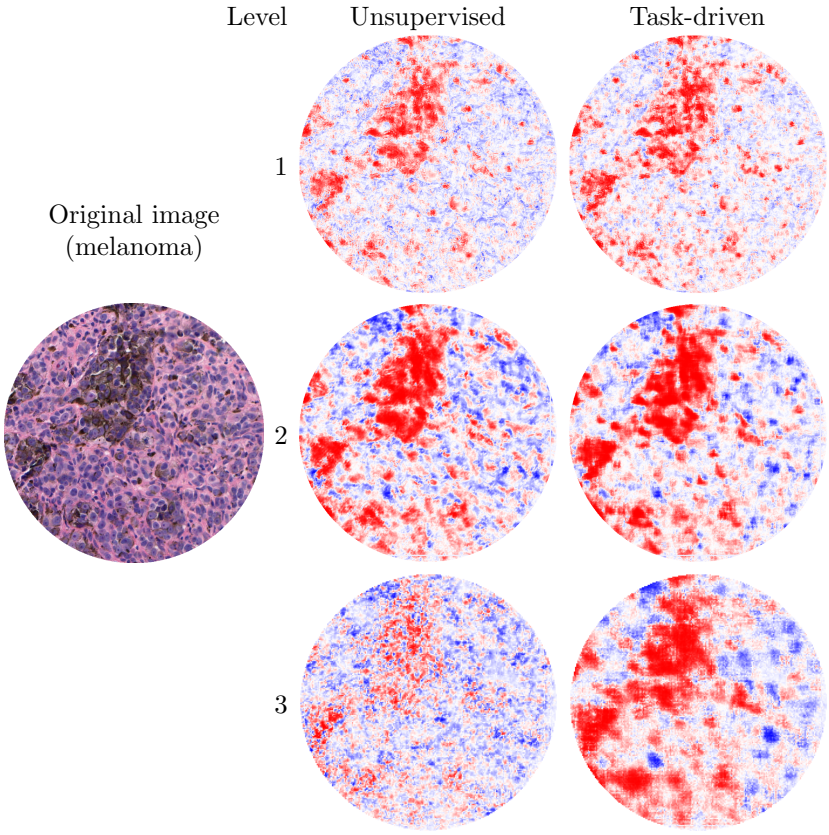

Figure 2.5: Relevance maps for a sample image: red indicates features associated with melanoma; blue indicates benign nevi.

2.6 Discussion

I have studied the performance of different representations in classifying histology images of melanoma and breast cancer. Hand-crafted features are difficult to develop and challenging to adapt to new data sets. Learned features - with dictionary learning or deep learning - produce superior results in classifying tumor tissue.

Deep learning results showed that transfer learning with a pre-trained CNN performed well even on the disparate image type of H&E histology. Further, the non-RGB transformation of extracting stain channels performed equally well, regardless of the channel ordering. This indicates that the power of pre-trained CNN features is not due to the individual features learned, but the space that they span. This conclusion is in line with Szegedy et al. who found that "there is no distinction between individual high level units and random linear combinations of high level units . . . suggest[ing] that it is the space, rather than the individual units, that contains the semantic information in the high layers of neural networks." [Szegedy et al., 2013] With unsupervised dictionary learning, stain normalization and the extraction of stain channels did provide a measurable increase in classification performance.

I have shown the application of hierarchical task-driven dictionary learning in predicting the diagnosis of melanoma and the subtype of breast tumors. The patch-level classification results indicate that the task-driven method has great promise in learning subtle features that distinguish classes. Chapters 3 and 4 will study multiple instance learning to handle the weak patient-level labels. Experiments in these chapters will show the importance of this step in training models and, in particular, Chapter 4 will show how it is critical during feature learning. Integrating some of these same techniques into task-driven dictionary learning could provide similar benefits.

CHAPTER 3: MULTIPLE INSTANCE LEARNING FOR HETEROGENEOUS IMAGES WITH AN SVM[1]

Automatic classification of histology images can be used to predict diagnosis, grade, or subtype. For diagnosis, the presence of even a small region of tumor indicates cancer. Tumors can also be grouped into subtypes, but the image may contain a heterogeneous mix of tissue types, making classification challenging as only part of the image is relevant. Tumors of different histologic types may also belong to the same genomic subtype, causing heterogeneous phenotypes within a subtype. In addition to classifying the whole sample, quantifying the subtype or other biomarker heterogeneity can provide an important measure for characterizing tumors [Hiley and Swanton, 2014; McGranahan and Swanton, 2015].

Multiple instance (MI) learning is commonly applied for diagnosis by breaking a large image or multiple images into smaller regions [Kandemir and Hamprecht, 2014; Xu et al., 2014a,b]. All data from a particular patient is referred to as a "bag" and each image region is called an "instance." A bag can have many instances. Labels of cancer and non-cancer are available at the patient or bag level, not the instance level, making this a weakly supervised learning problem. With the standard MI assumption, a sample is classified as positive if at least one of its instances is positive and negative otherwise. This asymmetric relationship works well for diagnosis, but not for problems such as subtype classification in which there is no distinctive "positive" and "negative" class. Diagnosis also requires that the presence of even a small region of tumor should produce a classification of tumor. For histologic and genomic subtypes, it is more appropriate to assign a label to an image based on the properties of multiple regions.

Motivated by the problem of tumor subtyping, I explore MI methods that can treat classes symmetrically and address tumor heterogeneity. These methods enable bag-level predictions as well as instance-level, providing a critical means for interpreting the classification model. Figure

[1]Some of the methods in this chapter were published in npj Breast Cancer [Couture et al., 2018b]. A more thorough statistical analysis of results from the journal paper was joint work with Lindsay Williams and is reproduced in Appendix A.

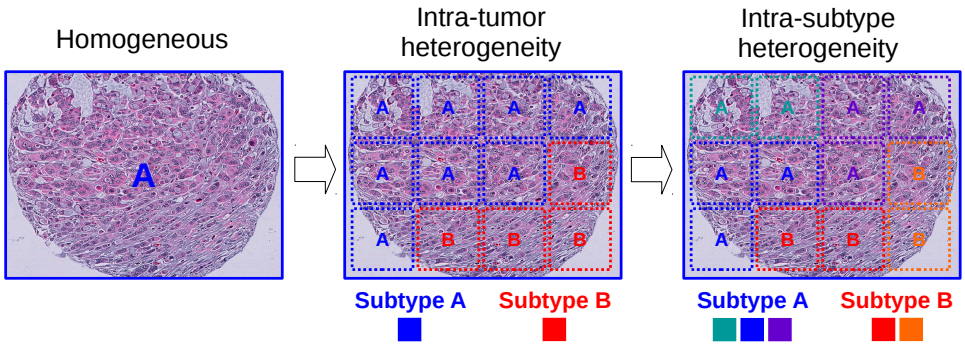

Figure 3.1: The proposed model handles intra-tumor heterogeneity by applying MI learning to multiple regions of an image using a feature set than can capture intra-subtype heterogeneity.

3.1 demonstrates the goals of this work in capturing intra-tumor and intra-subtype heterogeneity. Breaking a single image into multiple samples accommodates intra-tumor heterogeneity and also increases the amount of training data, which is critical when training data is limited in cancer research and many medical applications. Intra-subtype heterogeneity is handled with the use of an appropriate feature set that can capture the variety of appearances of a single class. Pre-trained CNN features are used in this chapter, although a different feature set, such as dictionary learning (Chapter 2), can also fit into this framework. Fine-tuning a CNN for the MI framework will be discussed in Chapter 4.2.

In this chapter, I examine two simple techniques for adapting MI learning to histology images: 1) selecting an appropriate function to aggregate instance predictions for histology classification (Section 3.2) and 2) generating optimally sized image regions from larger images (Section 3.3). I also propose two new methods to further the state of the art in MI learning: 1) an iterative method for learning latent instance labels under different MI assumptions (Section 3.4) and 2) aggregating instance predictions with a quantile function (Section 3.5). The implementation discussed in this chapter using an SVM can accommodate any type of image feature.

3.1 Related Work

Figure 3.2 provides an overview of related work, showing how the current work is set apart.

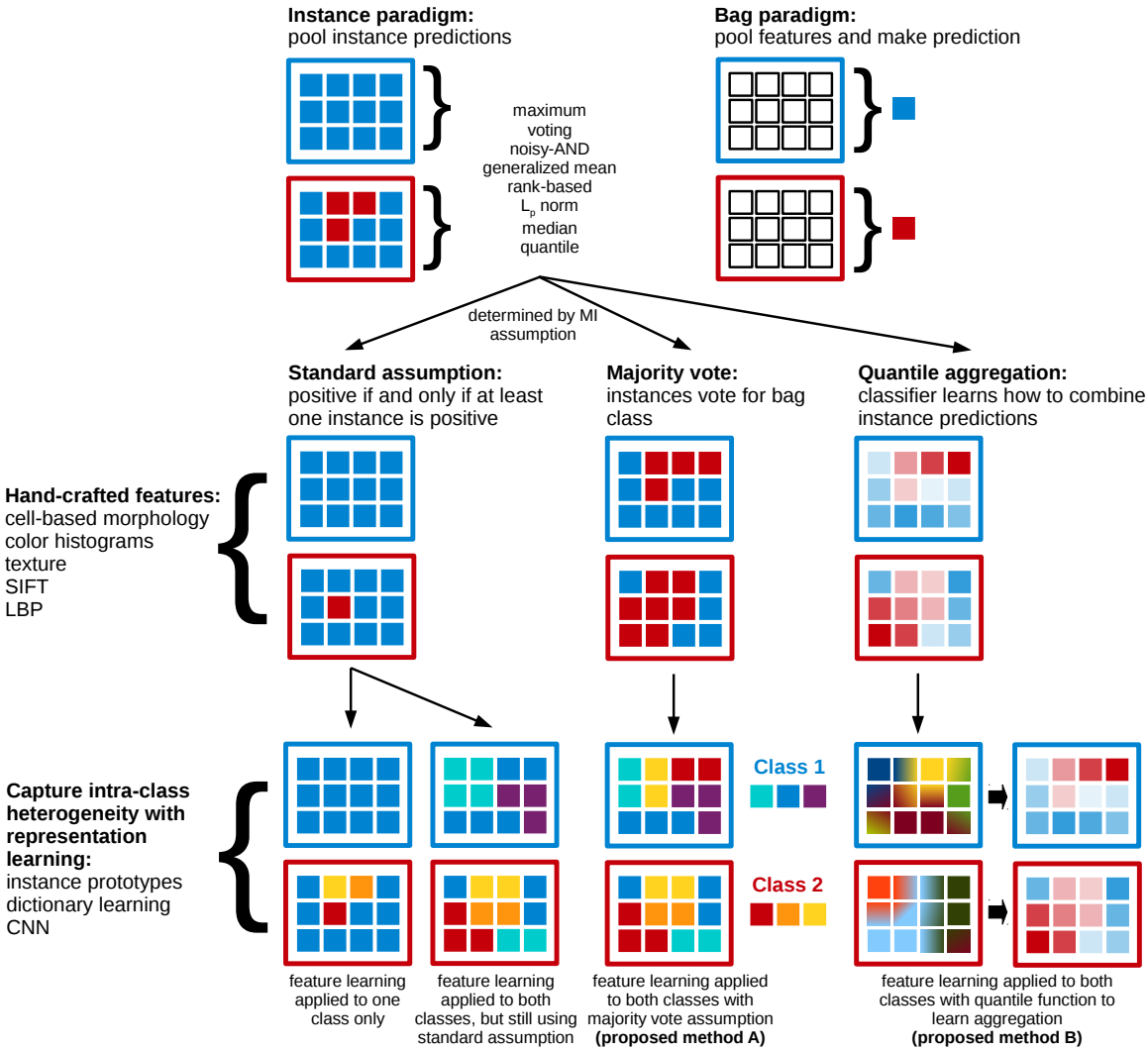

Figure 3.2: Instance paradigm methods first classify individual instances and then pool the predictions, while bag paradigm methods pool the instance features and make a single prediction. Instance paradigm methods also differ in whether they use the standard assumption positive if and only if at least one instance is positive), the majority vote, or learn how instance predictions should be aggregated, such as with the proposed method of quantile aggregation. If classes are homogeneous, a simple mean of hand-crafted cell features or patch descriptors is sufficient to characterize each image. When classes are heterogeneous, learning a representation becomes important. Proposed method A uses the majority vote instead of the standard assumption, along with feature learning and a iterative method to learn the latent instance labels. Proposed method B uses a quantile function of instance predictions to predict the bag class, thus capturing a wider range of possible MI assumptions.

Reduction to Single Instance Learning. Many image classification solutions for histology and other data types turn the problem into a fully supervised one by representing each bag as a single feature vector [Chen and Wang, 2004; Chen et al., 2006] or applying a specialized kernel [Zhou et al., 2009]. This class of methods can only make predictions at the bag level, not the instance level, so is not suitable for characterizing tissue heterogeneity or interpreting results. The methods that I present make use of an instance classifier.

Instance-level Methods. Rather than making decisions at the bag level, other MI approaches design classifiers to operate on individual instances and then aggregate their output scores or decisions. Andrews et al. developed mi-SVM, in which they apply an SVM to MI learning by iteratively learning the latent instance labels while enforcing the standard MI assumption [Andrews et al., 2002]. This class of score- or decision-level fusion methods is able to make use of a larger number of samples drawn from the set of instances when training the classifier. However, mi-SVM still follows the standard assumption: treating classes asymmetrically. I use the power of mi-SVM in learning latent instance labels but adapt it for a wider range of possible MI assumptions (Section 3.4).

MI Assumptions. For the standard MI assumption, all instances in a negative bag are negative and at least one instance in each positive bag is positive. This asymmetric definition treats positive bags differently than negative. MI techniques have been applied to histology for distinguishing images containing cancer from those that are cancer-free, using the standard assumption [Kandemir and Hamprecht, 2014; Xu et al., 2014a,b]. This asymmetric definition is not appropriate for classifying tumors by subtype. I propose a more general MI assumption in which a given percentage of instances must be positive and a method to learn the latent instance labels. Further, I propose the quantile function as a method for learning to aggregate instance predictions and for use when a suitable MI assumption for a particular task is unknown.

Bag- vs. Instance-level Predictions. Cheplygina et al. compare the stability of MI methods for three biomedical applications and find that the best bag-level classifier is not always the best instance-level classifier [Cheplygina et al., 2015]. mi-SVM is the most stable of the methods tested. Vanwinckelen et al. also compare instance-level and bag-level classification,

showing that the correlation varies widely by data set domain, learner assumptions, and performance measures [Vanwinckelen et al., 2016]. They also compared MI methods with their single instance (SI) counterpart, finding that often the SI method outperforms the MI algorithm. In some cases this may be due to a high witness rate - the proportion of positive instances in positive bags [Carbonneau et al., 2017]. To address this, Wang et al. optimize both the bag-level and instance-level loss by including them both in the cost function [Wang et al., 2015b]. My work addresses the challenge of producing an accurate instance- and bag-level classifier by better bridging the gap between them with a more powerful pooling function (Section 3.5).

Image Heterogeneity. Intra-class heterogeneity can be accounted for by forming multiple prototypes for each class; however, initial research in this direction focused on the standard assumption and only applied heterogeneity to the pathological case [Xu et al., 2014b; Li et al., 2015; Wang et al., 2013; Varol et al., 2015]. When each class represents a different subtype of the disease, heterogeneity should be accounted for in all classes. While dictionary learning applied to all classes could address this deficiency, existing methods still remain focused on the standard assumption [Shrivastava et al., 2015; Song et al., 2013; Jiao and Zare, 2015].

CNNs produce another powerful feature set for capturing heterogeneity. Hou et al. use a CNN with an iterative MI method for predicting cancer subtypes from whole slide histology images [Hou et al., 2016]. The standard MI assumption does not apply, so they must learn which image patches belong to the labeled class of the slide. Different methods of aggregating patch predictions were tested: maximum, voting, and a histogram of predictions with a logistic regression classifier. Their iterative MI method uses EM to maximize the data likelihood and does not take into consideration the MI assumption chosen; the iterative method that I propose learns the latent instance labels given the MI assumption as a constraint. My quantile aggregation method is also more suitable than a histogram of predictions because it easily accommodates a non-uniform distribution of predictions without needing to specify bin sizes. I experiment with pre-trained CNN features in this chapter and end-to-end training of a CNN in the following chapter.

3.2 Aggregation Functions for Single Instance Learning

In the following sections, a bag is represented by X_n $(n = 1, ..., N)$, has a label $Y_n \epsilon \{-1, 1\}$, and contains instances $x_{n,i}$ for $i = 1, ..., M_n$. The instances $x_{n,i}$ have unknown labels $y_{n,i} \epsilon \{-1, 1\}$. A classifier f predicts the class of individual instances, and a function g aggregates these instance scores $s_{i,n}$ into a bag score S_n:

$$\hat{y}_{n,i} = \text{sgn}(s_{n,i}) \quad s_{n,i} = f(x_{n,i})$$

$$\hat{Y}_n = \text{sgn}(S_n) \quad S_n = g(\{s_{n,1}, ..., s_{n,M_n}\}).$$

In the simplest form of MI learning, Single Instance Learning (SIL), all instances are given their bag label and are used to train an instance classifier f. Instance predictions can be aggregated into a bag prediction in different ways. For the standard MI assumption, taking the maximum of all instance predictions is appropriate:

$$g_{max}(\{s_{n,1}, ..., p_{n,M_n}\}) = \underset{i=1,...,M_n}{\text{argmax}}\ s_{n,i}.$$

For many of the classification tasks for histology discussed in this work, the median

$$g_{median}(\{s_{n,1}, ..., s_{n,M_n}\}) = \text{median}(\{s_{n,1}, ..., s_{n,M_n}\})$$

is more suitable. More generally, a given percentile q could be used:

$$g_{percentile}(\{s_{n,1}, ..., s_{n,M_n}\}) = \text{percentile}(\{s_{n,1}, ..., s_{n,M_n}\}, q).$$

3.3 Generating Instances

MI algorithms typically operate on either the instance or the bag paradigm. With the instance paradigm, class predictions are made for each instance and are aggregated to predict the class for the entire bag. Alternatively, the bag paradigm aggregates the features of instances and makes a single prediction for the entire bag. In the case of images, some middle ground can be found by controlling how each image (bag) is split into regions (instances). Larger image

regions ensure that, in the presence of heterogeneity, some of the labeled class is present in each region, enabling more accurate predictions at the instance level. Smaller image regions can result in more training data for the classifier, which is particularly important when labeled data is limited. They also enable predictions to be made on smaller image regions, creating a more interpretable result. The instance size is set by selecting the region size on which to compute features.

3.4 Iterative Multiple Instance Learning

Rather than assigning the bag label to each instance, I propose a method to learn the latent instance labels in order to improve the instance classifier. A particular MI assumption must be chosen corresponding to the aggregation functions discussed in Section 3.2. For the tasks I study in classifying histology images the classes are symmetric, so I chose the assumption that more than half of the instances in a bag must belong to the bag class. More generally, this could be that a given percentage of instances must belong to a particular class.

This MI formulation is based on the linear soft-margin SVM and is an extension of mi-SVM by Andrews et al. [Andrews et al., 2002]. I optimize jointly over the possible classifiers (w, b) and latent instance labels $y_{n,i}$ with an SVM:

$$\min_{\{y_{n,i}\},w,b,\{\xi_n\}} \frac{1}{2}||w||^2 + C\sum_{n=1}^{N}\sum_{i=1}^{M_n}\xi_{n,i} \tag{3.1}$$

such that $y_{n,i}(< w, x_{n,i} > +b) \geq 1 - \xi_{n,i}$ and $\xi_{n,i} \geq 0$ for all $n = 1, ..., N$ and $i = 1, ..., M_n$. In order to enforce the aggregation of instances into bags with the correct label, an additional constraint is needed. For the standard assumption with mi-SVM, this is

$$\sum_{i=1}^{M_n} \mathrm{I}(\hat{y}_{n,i} = 1) > 0 \text{ if and only if } Y_n = 1$$

where indicator function $\mathrm{I}(z)$ is 1 if z is true and 0 otherwise. I replace this with the requirement

that a given percentage of instances q must belong to the bag class, expressed as

$$\begin{aligned}\sum_{i=1}^{M_n} \mathrm{I}(\hat{y}_{n,i} = 1) &\geq \frac{q}{100} M_n \text{ for all } n \text{ such that } Y_n = 1 \text{ and} \\ \sum_{i=1}^{M_n} \mathrm{I}(\hat{y}_{n,i} = 1) &< \frac{q}{100} M_n \text{ for all } n \text{ such that } Y_n = -1.\end{aligned} \tag{3.2}$$

For histology, the median is chosen ($q = 50$).

I use an iterative method to jointly optimize over the possible classifiers and latent instance labels, as outlined in Figure 3.3. The instance labels $y_{n,i}$ are initialized as the class of the bag Y_n and are used to train an SVM. Predictions are then made for all instance labels $\hat{y}_{n,i}$ using the SVM. Instance labels must then be adjusted to meet the constraint (3.2). Instances within each bag are sorted according to the classifier output $p_{n,i}$. For max aggregation, the highest scoring instance in positive bags is set to +1, while all the instances in negative bags are set to -1. For median aggregation, the instances are first sorted, and then the highest scoring instances in positive bags are set to +1 until half of all instances are positive; the lowest scoring instances are set to -1 in negative bags. The same procedure can be applied for an arbitrary percentile q. These alternating steps are repeated until fewer than 0.1% of instances change label or some maximum number of iterations is reached. The label $\hat{Y}_n$ of a novel bag X_n can then be predicted as already described in Section 3.2.

3.5 Quantile Aggregation

Prior work has shown that assigning each instance the bag label during training (as in SIL, Section 3.2) can perform quite well in terms of both bag- and instance-level accuracy when the witness rate is fairly high [Carbonneau et al., 2017]; however, bag-paradigm methods still perform better than instance-paradigm methods on some data sets [Wang et al., 2018]. In order to preserve the benefit of an instance-level classifier in providing interpretability but still achieve a high bag-level accuracy, I propose a new method for aggregating instance predictions. This method will also aid applications in which the most suitable aggregator is unknown, as it can learn how much heterogeneity to accommodate.

I propose a method of quantile aggregation that predicts the bag class from the quantile

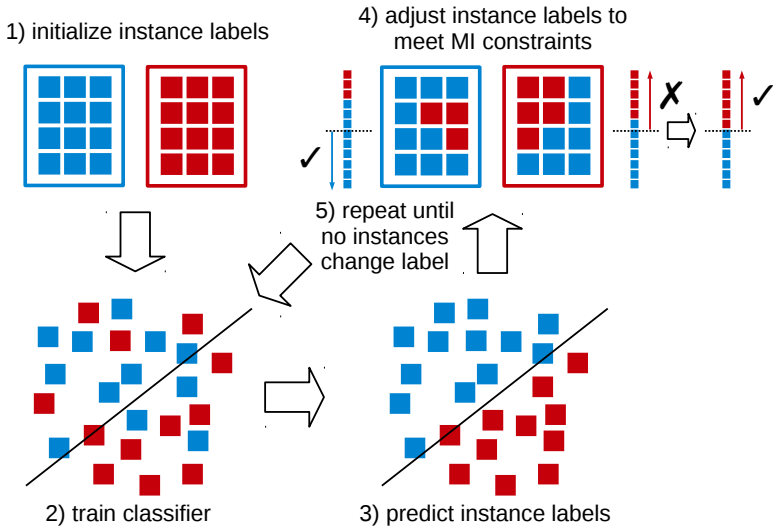

Figure 3.3: Overview of iterative MI method. 1) Instance labels are initialized according the the bag label. 2) An instance classifier is trained using these labels. 3) The classifier is used to predict the label of each instance. 4) Instance labels are adjusted until the MI constraints are met (e.g., for median aggregation, half of all instances must belong to the bag class). 5) This procedure is repeated until convergence.

function (QF) of instance predictions. Instance predictions are aggregated with the QF, and a bag classifier is trained to predict the bag class from the QF. The QF is the inverse cumulative distribution and represents the boundary points between fractions of the population [Broadhurst, 2008]. For random variable X the QF assigns to each probability p the value x for which $\Pr(X \leq x) = p$. An instance classifier f is first trained using the bag labels as instance labels. If the instance predictions for bag n are represented by $S_n = \{s_{n,1}, ..., s_{n,M_n}\}$, the q-th Q-quantile is the value z such that $\Pr(S_n \leq z) = (q - 0.5)/Q$. To form the QF, I first sort S_n into the set $\tilde{S_n} = \text{sorted}(S_n)$. The sorted values in $\tilde{S_n}$ are used to extract the QF vector as $Z_n = [z_{n,1}, ..., z_{n,Q}]$, where $z_{n,q} = \tilde{s}_{n,\lceil M_n(q-0.5)/Q \rceil}$. Another classifier g is trained on this bag-level feature set to predict the bag class:

$$\hat{Y}_n = \text{sgn}(S_n) = \text{sgn}\left[g(Z_n)\right].$$

The data for fitting the instance classifier and bag classifier must be disjoint. The training bags are randomly split into K folds and K instance classifiers are trained, each using all but one of the folds of data. Predictions are made using the remaining fold of data and used to compute

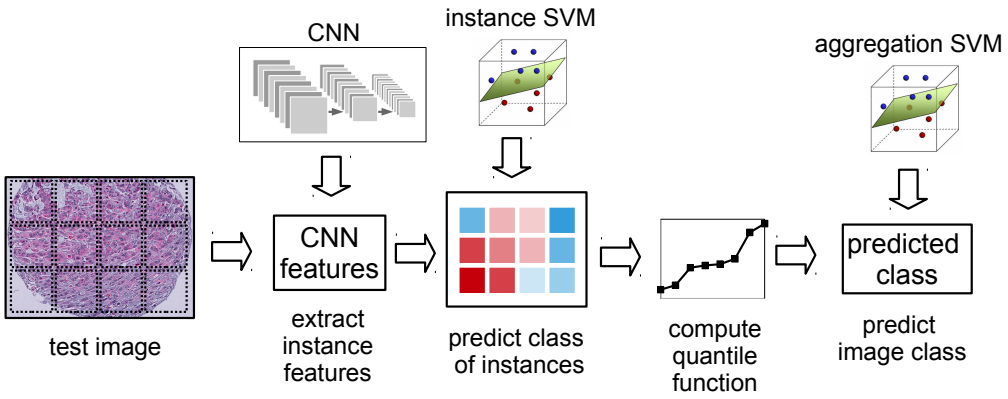

Figure 3.4: At test time, a pre-trained CNN is used to extract features from the image, with a local mean pool to produce each instance. Instance predictions are made with an SVM and a quantile function is used to summarize these predictions over each image. The aggregation SVM then uses the computed quantile function to predict the class of the image.

the QF for each bag. The computed QFs are then used to train the aggregating classifier. This ensemble method ensures that the instance predictions used to train the bag classifier are not biased by being used in training the instance classifier. At test time, the mean of the ensemble predictions on instances is used to form the QF. Figure 3.4 provides an overview of the steps taken to predict the class of a novel image.

The QF captures how much of each class is present in an image, so this technique enables the aggregating classifier to learn how much heterogeneity to expect in images of each class. It can also learn whether the median or some higher or lower quantile is a strong predictor, thus encapsulating many other possible aggregating functions. The QF is more suitable than a histogram as the bin sizes do not need to be specified. It provides a better discretization than a histogram and easily accommodates a non-uniform distribution of predictions from the instance-level SVM. Euclidean distance can be then applied to the QF [Broadhurst, 2008], making the linear kernel still suitable for the bag-level classifier.

3.6 Sample Weighting by Class

When class labels are imbalanced, it is common practice to weight training samples inversely proportional to the frequency of their class label during classifier training. That is, samples of class c are weighting by

$$w_c = \frac{N}{CN_c}$$

where N is the number of samples, N_c is the number of samples of class c, and C is the number of classes. However, there can be further imbalances within a class due to imbalanced data of a secondary classification. In the context of breast tumor histology, this occurs with ER status and intrinsic subtype, in which there are few samples for ER positive/high grade and non-Basal/high grade, leading to poor classification accuracy of high grade tumors. To improve classification for the secondary class g, I weight samples of class c/class g by

$$w_{c,g} = \frac{N}{CGN_{c,g}}$$

where $N_{c,g}$ is the number of samples of class c/class g and G is the number of classes in this secondary classification.

3.7 Experiments

To validate these MI methods, I examined the classification accuracy on two different data sets of breast tumor H&E histology images.

3.7.1 Data Sets

BreaKHis. Samples of benign and malignant breast tumors were stained with H&E and imaged at 200x visual magnification/20x objective lens. These samples came from 82 different patients, with a total of 2013 images at this particular magnification. Previous results have been reported on this data set, so I used the same training and testing procedure [Spanhol et al., 2016b,a, 2017]. The data set has been divided into five different random splits consisting of 70% training and 30% testing images. The recognition rate was measured for each, and then the mean was taken over the five splits. The recognition rate is defined as

$$\text{Recognition rate} = \frac{\sum_p \text{Score}_p}{\text{Total number of patients}}$$

with a patient score of

$$\text{Score}_p = \frac{N_{rec}}{N_p}$$

where N_{rec} is the number of images correctly classified and N_p is the number of images for the patient.

CBCS. This data set consists of 1713 patient samples from the Carolina Breast Cancer Study, Phase 3 [Troester et al., 2018]. There are typically four cores per patient (5970 cores total), with each core image having a diameter of around 2400 pixels. I used 5-fold cross validation and computed the average across folds. Classification accuracy and AUC were measured for Basal vs. non-Basal intrinsic subtype, ER positive vs. negative, and grade 1 vs. 3.

3.7.2 Implementation Details

Image Pre-processing. For the CBCS data set, images were first color and intensity normalized to standardize the appearance across slides, countering effects due to different stain amounts and protocols, as well as slide fading. I used the method by Niethammer et al. that estimates the stain vectors for hematoxylin and eosin and with that normalizes each image [Niethammer et al., 2010]. The hematoxylin, eosin, and residual channels were extracted from the normalization process and used as input for feature extraction. No pre-processing was done for the BreaKHis data set.

Features. The experiments that follow use features from the pre-trained CNN AlexNet [Krizhevsky et al., 2012]. The CNN was applied in a fully convolutional manner using the output from the fourth convolutional layer. Other layer outputs were tested as well, with the best results achieved from the fourth layer. Further experiments later in this section applied the pre-trained CNN VGG16 and used the output from the fourth set of convolutional layers [Simonyan and Zisserman, 2015].

Software. The methods in this chapter were implemented in Python, using Keras to extract pre-trained CNN features and scikit-learn for SVM classification. SVM hyperparameters were learned by grid search using five-fold cross-validation on the training set.

Hyperparameters. The SVM hyperparameter C was selected by grid search with five-fold cross-validation on the training set. For the quantile aggregation methods, 16 quantiles were

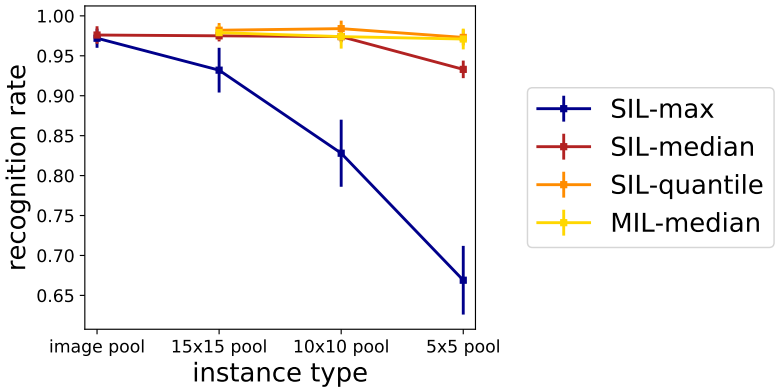

Figure 3.5: Recognition rate for different MI methods on the BreaKHis data set.

used, except for the case of image-level pooling in which case 4 quantiles were used.

3.7.3 Classification Results

Mean pooling was applied to the output from the fourth convolutional layer of AlexNet using different pool sizes to create instances from each image. I experimented with different pool sizes on each data set, using the methods Single Instance Learning with max (SIL-max), median (SIL-median), and quantile (SIL-quantile) aggregation as well as iterative Multiple Instance Learning with median aggregation (MIL-median).

Figure 3.5 shows a comparison of the recognition rate results on the BreaKHis data set. SIL-max clearly performed poorly, especially for smaller pooling sizes where there were many instances. While the recognition rate of SIL-median started to drop off as the number of instances increased, SIL-quantile and MIL-median maintained a close to consistent recognition rate. The previous best image-level recognition rate reported on this data set was 86.3%, which also used features from a pre-trained CNN to train a classifier [Spanhol et al., 2017]; my best result was 98.4% using SIL-quantile. Other slightly poorer results on this data set tried fine-tuning a small CNN [Spanhol et al., 2016b] and an SVM with hand-engineered features [Spanhol et al., 2016a].

Results on the CBCS data set are shown in Figure 3.6, with both the AUC and classification accuracy measured. Once again, SIL-max did not hold up when there were many instances. SIL-median performed well with few instances but drooped for larger numbers of instances. SIL-quantile and MIL-median remained fairly steady as the number of instances increased, with

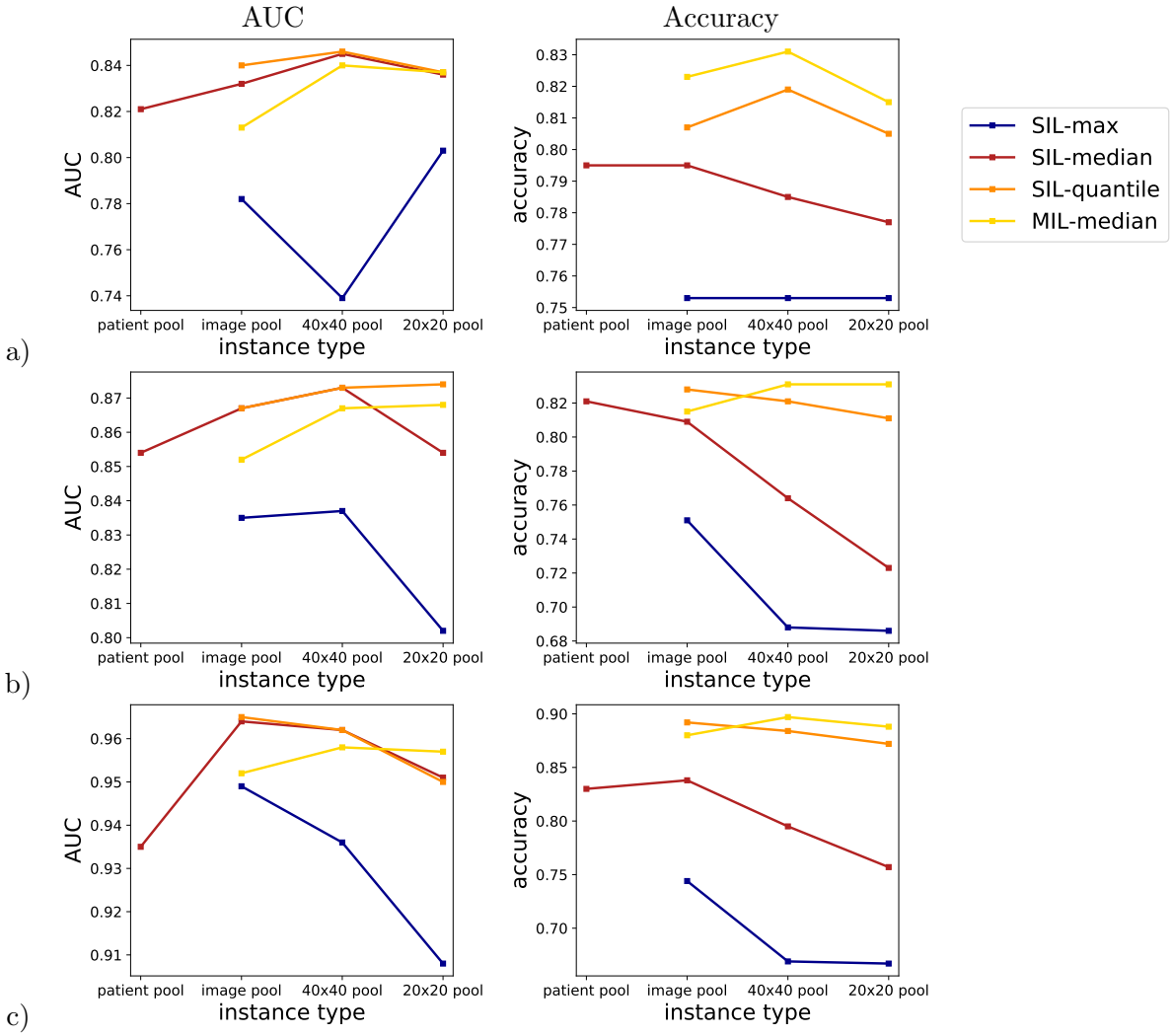

Figure 3.6: AUC and classification accuracy for different MI methods on the CBCS data set: a) Basal vs. non-Basal, b) ER positive vs. negative, c) grade 1 vs. 3.

their performance sometimes even increasing when more instances were used. The graphs in Figure 3.6 also show a difference between the results for AUC and classification accuracy. While the AUC of SIL-quantile and MIL-median is similar to that of SIL-median for fewer instances, an increase in classification accuracy was observed for this same situation; SIL-quantile and MIL-median produced a large increase in classification accuracy for Basal vs. non-Basal and grade 1 vs. 3.

Further improvements in classification accuracy can be obtained with a more powerful CNN such as VGG16. I compare the results in Figure 3.6 using AlexNet with those achieved with VGG16 in Table 3.1. The results are shown for each MI method using the pooling (instance

Method	Basal vs. non-Basal		ER status		Grade 1 vs. 3	
	AlexNet	VGG16	AlexNet	VGG16	AlexNet	VGG16
SIL-median	0.795	0.805	0.821	0.827	0.838	0.882
SIL-quantile	0.819	0.841	0.828	0.857	0.892	0.907
MIL-median	0.831	0.851	0.831	0.874	0.897	0.923

Table 3.1: Classification accuracy using features from AlexNet or VGG16 for different MI methods. The best pooling (instance creation) strategy is selected for each method.

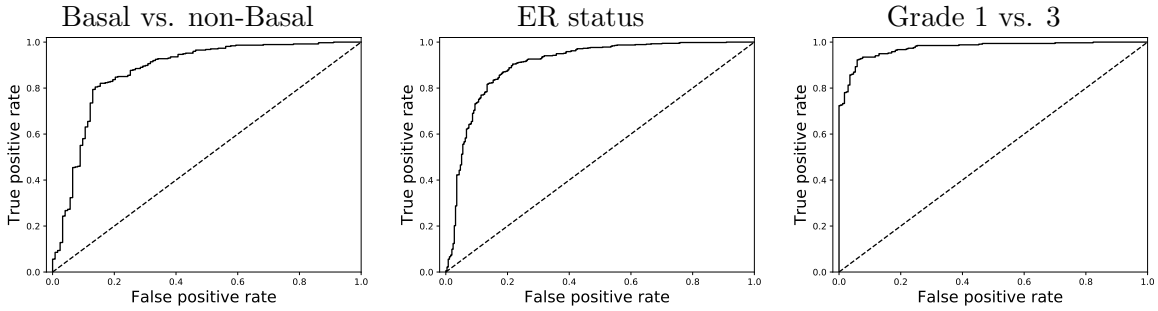

Figure 3.7: ROC plots for MIL-median with VGG16.

creation) strategy that produced the best result. By selecting the best results across pooling types, we can see once again that SIL-quantile consistently produced better results than SIL-median. Further, we see that MIL-median was the top performer overall. VGG16 provided up to a 4.4% improvement in classification accuracy. ROC plots for MIL-median with VGG16 are shown in Figure 3.7. Fine-tuning the CNN could produce a further increase in classification performance, which will be studied in the next chapter.

3.7.4 Sample Weighting by Class

I studied the use of sample weighting by grade in predicting ER status and intrinsic subtype. The non-grade-weighted method using SIL-quantile simply weights the samples inversely proportional to the number of samples in each class. The grade-weighted method also weights by low-intermediate vs. high grade, using the method in Section 3.6. This study was done using a subset of the CBCS data set, including 1203 samples for training and 401 for testing, with no cross-validation. Table 3.2 shows that the classification accuracy of high grade samples improved for both ER status and Basal vs. non-Basal, although the accuracy for low-int grade samples decreased somewhat.

	Overall	Low-int grade			High grade		
	accuracy	Sens.	Spec.	Acc.	Sens.	Spec.	Acc.
ER status							
Not grade weighted	0.835	1.00	0.273	0.938	0.850	0.613	0.706
Grade weighted	0.852	0.897	0.636	0.875	0.800	0.839	0.824
Basal vs. non-Basal							
Not grade weighted	0.785	1.00	0.00	0.949	0.935	0.302	0.629
Grade weighted	0.812	0.960	0.750	0.924	0.761	0.605	0.685

Table 3.2: Accuracy, sensitivity and specificity for ER status and Basal vs. non-Basal with and without weighting samples by grade.

3.7.5 Statistical Validation

A more thorough validation of the SIL-quantile method on a subset of the CBCS data set is provided in Appendix A. This work was done jointly with Lindsay Williams and was published in npj Breast Cancer [Couture et al., 2018b]. It studies the effectiveness of the SIL-quantile method in predicting five breast tumor histology properties: grade, ER status, intrinsic subtype, histologic subtype, and risk of recurrence.

3.7.6 Visualization of Heterogeneity

Instance-based MI methods like those studied here provide a means for interpretation and insight into class heterogeneity. I examined the class predictions across cores from the same patient and within each core. Figure 3.8 shows four cores from a single patient, along with the class predictions over different regions of the image that were generated from overlapping 400×400 pixel instances across the image. While three cores were predicted ER negative and Basal intrinsic subtype, the fourth was predicted mostly ER negative and non-Basal, indicating that some intra-tumor heterogeneity might be present across cores.

3.8 Discussion

The results in this chapter confirm that the standard MI assumption that takes the maximum across all instances in a bag is not appropriate for the tasks addressed in this work; median aggregation is more appropriate. The results also confirm that splitting images into multiple instances can outperform a single instance per patient. However, dividing images into

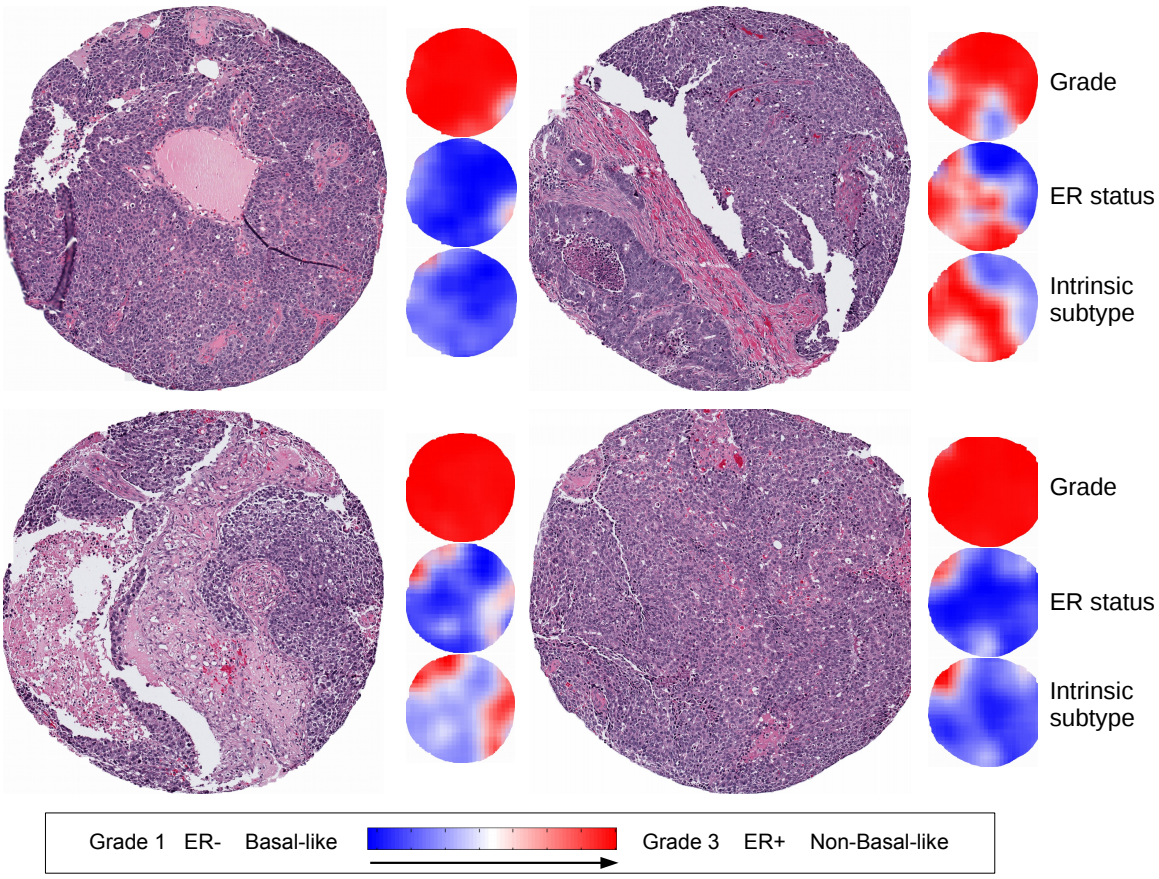

Figure 3.8: Four H&E cores from a single patient and heat maps indicating the class predictions from overlapping instances. Class probabilities are indicated by the intensity of red/blue color with greater intensity for higher probabilities. Uncertainty in the prediction is indicated by white. This patient was labeled as high grade, ER negative, and Basal intrinsic subtype.

too many instances can degrade performance for some MI methods. Quantile aggregation is one simple technique to maintain a high classification accuracy when there are many instances. This method aggregates instance predictions with the quantile function and learns how to combine them through a bag-level classifier. When an iterative method that learns the latent instance labels is paired with an appropriate aggregation function for the task (median for the histology tasks that I experiment on), a further increase in performance is observed but at the expense of a longer computation time. Prior work [Vanwinckelen et al., 2016; Wang et al., 2018] has shown that when the witness rate (the proportion of positive instances in positive bags) is high, MIL methods typically do not outperform SIL methods. This work shows a clear benefit from MIL in classifying histology.

The power of these SVM-based MI methods was demonstrated on four different tasks on H&E histology: benign vs. malignant tumor, grade 1 vs. 3, ER positive vs. negative, and intrinsic subtype Basal vs. non-Basal. The gold standard for the first two tasks is with pathologist review, while ER status is evaluated from immunohistochemical staining and intrinsic subtype from molecular methods. The latter two were previously not known to be predictable from H&E histology alone. By focusing on class predictions for local regions across the image, these methods provide much needed insight into tumor heterogeneity.

CHAPTER 4: MULTIPLE INSTANCE LEARNING FOR HETEROGENEOUS IMAGES WITH A CNN[1]

Deep learning has become the standard solution for classification when a large set of images with detailed annotations is available for training. When the annotations are weaker, such as with large, heterogeneous images, we turn to multiple instance (MI) learning. The image (called a bag) is broken into smaller regions (called instances). We are given a label for each bag, but the instance labels are unknown. Some form of pooling aggregates instances into a bag-level classification. By integrating MI learning into a convolutional neural network (CNN), one can learn an instance classifier and aggregate the predictions so the entire system is trained end-to-end [Kraus et al., 2016; Sun et al., 2016; Jia et al., 2017].

I propose a more general approach for aggregating instance predictions that looks at the full distribution by pooling with the quantile function (QF) and learning how much heterogeneity to expect for each class. As data augmentation is especially critical in training large CNNs, I also created an augmentation technique for training MI methods with a CNN (Fig. 4.1). Through MI augmentation, I study the importance of the MI formulation during training.

Using MI learning to make class predictions over smaller regions of the image provides insight into how different parts of the image contribute to the classification. Visualizing the instance predictions provides a method of interpretability that I demonstrate on a data set of breast tumor tissue microarray (TMA) images stained with H&E by predicting grade, receptor status, intrinsic subtype, histologic subtype, and risk of recurrence. Some of these tasks are not previously known to be achievable from H&E alone. My quantitative results conclude that the MI component is critical to successful classification, demonstrating the importance of accounting for heterogeneity. This method could provide future insights into tumor heterogeneity and its connection with cancer progression [Hiley and Swanton, 2014; McGranahan and Swanton, 2015].

While Chapter 3 focused on SVM-based methods than can handle any type of feature set,

[1]Most of the work in this chapter was presented at the International Conference on Medical Image Computing & Computer Assisted Intervention in 2018 [Couture et al., 2018a].

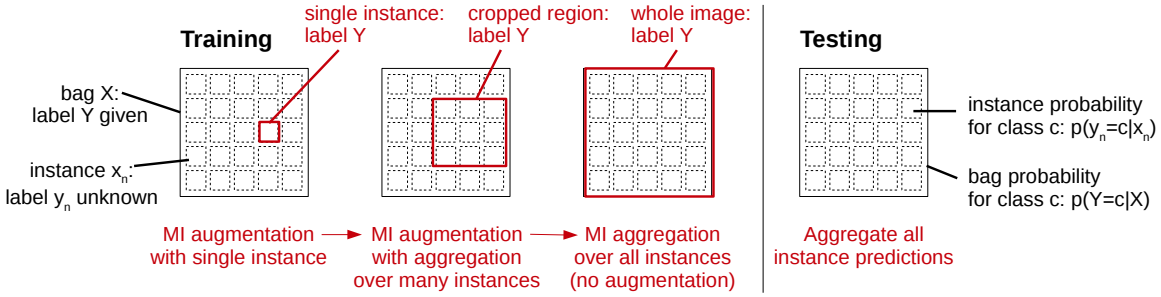

Figure 4.1: In MI learning, each bag contains one or more instances. Labels are given for the bag but not the instances. MI augmentation is a technique to provide additional training samples by randomly selecting a cropped image region and the instances within it. When the bag label is applied to a small number of instances, it is weak because this small region may not be representative of the bag class. Applying the bag label to larger cropped regions provides a stronger label, while still providing benefit from image augmentation. Training with the whole image maximizes the opportunity for MI learning but restricts the benefits of image augmentation. *At test time, the whole image is processed and the predictions from all instances are aggregated into a bag prediction.*

this chapter addresses MI learning with a CNN, enabling end-to-end training. Some background on the two main contributions of this chapter is discussed in Section 4.1. Section 4.2 sets up the CNN-based MI framework by adding an MI layer. The QF aggregation method (first introduced in Chapter 3) is presented as a CNN layer in Section 4.3 and MI augmentation in Section 4.4. Section 4.5 provides quantitative validation of the methods presented and demonstrates insight into tumor heterogeneity.

4.1 Background

Aggregating Instance Predictions. A *permutation invariant* pooling of instances is needed to accommodate images of different sizes, whereas a fully connected neural network cannot accommodate images of different sizes. Existing pooling approaches are very aggressive; they compute a single number rather than looking at the distribution of instance predictions. Most MI applications use the maximum, which works well for problems such as cancer diagnosis where, if there is a small amount of tumor, the sample is labeled as cancerous [Kandemir and Hamprecht, 2014; Xu et al., 2014b]. A smooth approximation, such as the generalized mean or noisy-OR, provide better convergence in a CNN [Kraus et al., 2016; Sun et al., 2016; Jia et al., 2017]. For other tasks, a majority vote, median, or mean is more appropriate. I include more

of the distribution by pooling with the QF and learning a mapping to the bag class prediction, improving the classification accuracy. My proposed method of quantile aggregation learns how to predict the bag class from instance predictions and thus could provide a solution when the most suitable aggregator is unknown. While Chapter 3 presented the QF integrated with an SVM-based classification method, this chapter adapts it for use in a CNN trained end-to-end. The QF is a new general type of feature pooling that could provide an alternative to max pooling in a CNN.

Training MI Methods with a CNN. Image augmentation is commonly applied in training a CNN by randomly cropping large portions of each image during each epoch. At test time the whole image is used. I propose MI augmentation, in which a subset of instances is randomly selected from each bag during each epoch of training. Instances are always the same size, but I choose how many instances to aggregate over. In selecting the number of instances, there are two extremes: a single instance vs. the whole bag. In the former, the bag label is assigned to each instance and is often called single instance learning. In the latter, MI aggregation is incorporated while training the bag classifier as in other MI methods [Andrews et al., 2002; Hou et al., 2016]. Comparison studies have found little or no improvement from these MI methods on some data sets [Vanwinckelen et al., 2016; Wang et al., 2018]. I found MI learning to be very beneficial and show that it is critical in dealing with heterogeneous data.

4.2 Multiple Instance Learning with a CNN

I denote a bag by X, its label by $Y \epsilon \{1, 2, ..., C\}$, and the instances it contains by x_n for $n = 1, ..., N$. The instance labels y_n are unknown. On a novel sample, an instance classifier f^c_{inst} predicts the probability of each class c as $s_{n,c}$ and a function f^c_{agg} aggregates these instance probabilities into a bag probability S_c:

$$s_{n,c} = f^c_{inst}(x_n) = \Pr(y_n = c|x_n) \qquad S_c = f^c_{agg}(s_{1,1}, ..., s_{N,C}) = \Pr(Y = c|X) \ .$$

MI learning can be implemented with many different types of classifiers [Andrews et al., 2002; Kandemir and Hamprecht, 2014; Vanwinckelen et al., 2016]. When implemented as a CNN, a

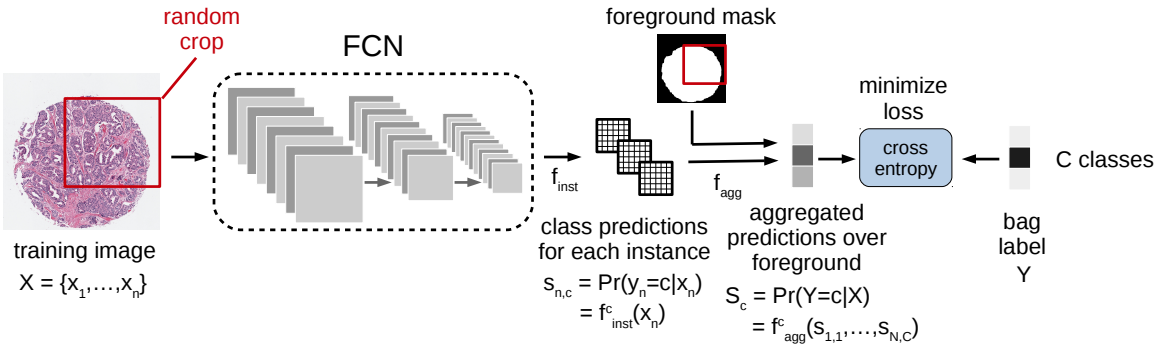

Figure 4.2: During training, a cropped region of a given size is randomly selected. An FCN is applied to predict the class, producing a grid of instance predictions. The instance predictions are aggregated over the foreground of the image (as indicated by the foreground mask) using quantile aggregation to predict the class of the cropped image region. With a cross-entropy loss applied, backpropagation then learns the FCN and aggregation function weights. At test time, the whole image is used.

fully convolutional network (FCN) forms the instance classifier f_{inst}, followed by a global MI layer for instance aggregation f_{agg}. The FCN consists of convolutional and pooling layers that downsize the representation, followed by a softmax operation to predict the probability for each class. For an input image of size $w \times w \times 3$, the FCN output is downsized to $w_d \times w_d \times C$. An instance is defined as the receptive field from the original image used in creating a point in this $w_d \times w_d$ grid; the instances are overlapping. The MI aggregation layer takes the instance probabilities and the foreground mask for the input image (downscaled to $w_d \times w_d$), thereby aggregating over only the foreground instances. Figure 4.2 provides an overview.

4.3 Multiple Instance Aggregation

Instance predictions can be used to form a bag prediction in different ways. The bag prediction function should be invariant to the number and spatial arrangement of instances, so some pooling of predictions is needed. Mean aggregation is well suited for global pooling as it is permutation invariant and can incorporate a foreground mask for the input image. Denoting the value for each element in the mask as $m_n \epsilon \{0, 1\}$, the mean aggregation function is

$$S_c = f^c_{\text{mean agg}}(s_{1,1}, ..., s_{N,C}) = \frac{\sum_{n=1}^{N} m_n s_{n,c}}{\sum_{n=1}^{N} m_n}.$$

Mean pooling incorporates predictions from all instances, but a lot of information is lost

in compressing to a single number. I propose quantile aggregation to provide a more complete description of the instance predictions in a bag. The QF was first introduced in Section 3.5 of the previous chapter, where it was used with binary predictions; here it is used in a multi-class setup, requiring a separate QF for each class. If the instance predictions for class c are represented by $S_c = \{s_{1,c}, ..., s_{N,c}\}$, then the q-th Q-quantile is the value z such that $\Pr(S_c \leq z) = (q - 0.5)/Q$. To pool with the QF, I first sort S_c and exclude instances not in the foreground, leaving the set $\tilde{S}_c = \{\tilde{s}_{1,c}, ..., \tilde{s}_{\tilde{N},c}\}$. The sorted values in $\tilde{S}_c$ are used to extract the QF vector for each class c as $z_c = [z_{1,c}, ..., z_{Q,c}]$ where $z_{q,c} = \tilde{s}_{\lceil \tilde{N}(q-0.5)/Q) \rceil, c}$. The QF vectors for all classes are concatenated as $Z = [z_1, ..., z_C]$. I then use a softmax function operating on Z to predict the bag class. The QF from all classes is used in order to learn the interaction of different subtypes in a bag. Backpropagation through the QF for gradient descent optimization operates in a similar manner to max pooling by passing the gradient back to the instance that achieved each quantile.

4.4 Training with Multiple Instance Augmentation

Image augmentation by random cropping is an important technique for creating extra training samples that helps to reduce overfitting. I propose an augmentation strategy for MI methods to increase the number of training samples by randomly selecting a different subset of instances for each epoch. Figure 4.1 provides a visual description of this technique. I select a particular crop size for training and randomly crop the image to select the set of instances, such that each crop contains at least 75% foreground according to the foreground mask. It is important to note that the image is never resized and the instance size remains constant. For each crop size chosen, the FCN is applied to the cropped image at full resolution. MI augmentation is a strategy used during training. *As the MI aggregation layer is invariant to input size, the entire image and all its instances are always used at test time.*

4.5 Experiments

4.5.1 Data Set

My data set consists of 1713 patient samples from the Carolina Breast Cancer Study, Phase 3 [Troester et al., 2018]. There are typically four 1.0 mm cores per patient in the TMA, with a total of 5970 cores. Each core was selected from the H&E-stained whole slide by a pathologist such that it contains a substantial amount of tumor tissue. Each image has a diameter of around 2400 pixels and a maximum of 3500 pixels. One sample core is shown in Fig. 4.2. I used a random subset of half the patients for training and the other half for testing. Classification accuracy was measured for five different tasks, some of them multi-class: 1) histologic subtype (ductal or lobular), 2) estrogen receptor (ER) status (positive or negative), 3) grade (1, 2, or 3), 4) risk of recurrence score (ROR) (low, intermediate, or high), 5) genomic subtype (basal, luminal A, luminal B, HER2, or normal-like). Ground truth for histologic subtype and grade are from a pathologist looking at the original whole slide. ER status was determined from immunohistochemistry, genomic subtype from the PAM50 array [Parker et al., 2009], and ROR from the ROR-PT score-based method [Parker et al., 2009].

4.5.2 Implementation Details

The TMA images were intensity-normalized to standardize the appearance across slides [Niethammer et al., 2010]. The hematoxylin, eosin, and residual channels were extracted from the normalization process and used as the three-channel input for the rest of my algorithm. A binary mask distinguishing tissue from background was also provided as input.

I used the pre-trained CNN AlexNet [Krizhevsky et al., 2012] and fine-tuned with the MI architecture shown in Fig. 4.2. All five tasks were equally weighted in a multi-task CNN as shared features help to reduce overfitting. For each patient, ground truth labels were available for most tasks. The cross-entropy loss was adjusted to ignore patients missing a label for a particular task.

CNNs were fine-tuned using Keras with the Theano backend and the Adam optimizer. For mean aggregation, I ran 50 epochs with a learning rate of 10^{-4}. I then dropped the learning

rate to 10^{-5} and ran 10 more epochs and a further 10 epochs with a learning rate of 10^{-6}. For quantile aggregation, I held the pre-trained CNN fixed to begin with and ran 30 epochs with a learning rate of 10^{-3} and 10 epochs at 10^{-4} to train only the QF aggregation. I then ran the same training procedure as for mean aggregation to fine-tune the whole CNN with QF aggregation.

In addition to MI augmentation, I randomly mirrored and rotated each training image. To accommodate the larger cropped image sizes in GPU memory, I reduced the batch size. A typical image with tissue of diameter 2400 pixels produced a 68×68 grid of instances. After applying the foreground mask, there were roughly 3600 instances. $Q = 15$ quantiles were used in all experiments. There are typically four core images per patient; I assigned the patient label to each during training and, at test time, took the mean prediction across the images. Further MI learning could be done to address the multiple core images per patient; however, my current focus is only on MI learning within each image.

4.5.3 MI Augmentation and the Importance of MI Learning

I studied the effect of MI learning on large images by selecting the cropped image size for training. The smallest possible size is 227×227 (the input size for AlexNet), consisting of a single instance. When the bag label is applied to each instance during training, this is called single instance learning. Alternatively, a larger cropped region of size $w \times w$ can be selected; I tested multiples of 500 up to 3500 and used mean aggregation in this experiment. By assigning the bag label to this larger cropped region during training and keeping the instance size constant, I performed MI learning. Multiple random crops were obtained from each training image such that roughly the same number of pixels was sampled for each crop size (i.e., the whole image for the largest crop size of 3500, $\frac{3500^2}{w^2}$ random crops for a training crop of size w). For the largest crop size, the whole image was used without MI augmentation. Random mirroring and rotations were used for augmentation at all crop sizes. At test time the whole image was always used, with the bag prediction formed by aggregating across all instances.

Figure 4.3 shows that larger crop sizes *for training* significantly increased classification accuracy ($p < 10^{-3}$ with McNemar's test for w=500 vs. w=1500 on all tasks). The benefits

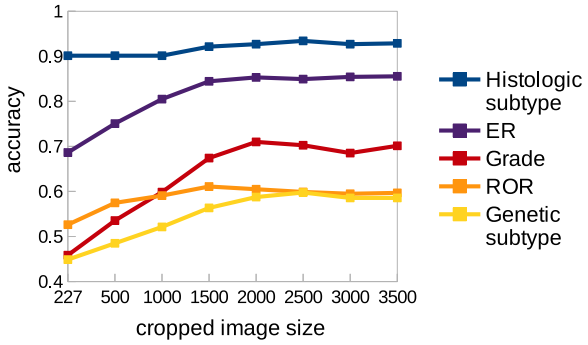

Figure 4.3: Classification accuracy using mean aggregation as the number of instances (cropped image size) used for training is increased, while keeping instance size constant.

leveled off for larger crops. As GPU memory requirements increase for larger crop sizes, selecting an intermediate crop size provides most of the benefits of MI augmentation.

Although it should not be surprising that a larger crop size at training works better, the magnitude of improvement was very significant. If the images were homogeneous (at the scale of a single instance, $w = 227$) then applying the bag label to each instance should produce a classification accuracy similar to when MI aggregation over the whole image is used during training. This is clearly not the case in Fig. 4.3. For example, ER status increases from 68.6% to 85.6% when MI learning was applied over the whole image. *This demonstrates the importance of MI learning and the effect of heterogeneity.* My data set consists of cores selected from a whole slide by a pathologist. MI learning may be even more crucial when classifying larger and more heterogeneous images like whole slides.

4.5.4 MI Aggregation

I compared aggregation methods by training my model on a crop size $w = 2000$ and taking the average classification accuracy over four runs. Table 4.1 shows that mean and quantile aggregation both significantly outperformed max aggregation ($p < 10^{-8}$ with McNemar's test). While quantile aggregation performance was similar to mean for some tasks, a significant increase in performance (93.1% to 95.2%) was observed for predicting the histologic subtype of ductal vs. lobular ($p < 10^{-10}$ with McNemar's test). This improvement was due to quantile aggregation predicting the bag class from a more complete view of the instance predictions using QF pooling, thereby capturing the heterogeneity.

Task	Max	Mean	Quantile
Histologic subtype	0.898 (0.004)	0.931 (0.004)	0.952 (0.003)
ER	0.683 (0.006)	0.833 (0.008)	0.841 (0.006)
Grade	0.408 (0.019)	0.680 (0.003)	0.676 (0.006)
ROR	0.542 (0.010)	0.595 (0.003)	0.582 (0.008)
Genomic subtype	0.321 (0.032)	0.548 (0.006)	0.544 (0.003)

Table 4.1: Average classification accuracy for different types of MI aggregation. The standard error is in brackets.

CNN architecture	Histologic subtype	ER	Grade	ROR	Genomic subtype
AlexNet [Krizhevsky et al., 2012]	0.910	0.822	0.663	0.563	0.521
VGG16 [Simonyan and Zisserman, 2015]	0.951	0.879	0.680	0.603	0.565
ResNet [He et al., 2016]	0.909	0.853	0.673	0.567	0.567
Inception-v3 [Szegedy et al., 2016]	0.910	0.862	0.694	0.581	0.563
Inception-Resnet-v2 [Szegedy et al., 2017]	0.901	0.869	0.700	0.577	0.573

Table 4.2: Classification accuracy for different CNN architectures using the mean aggregation method.

4.5.5 CNN Architecture

The results presented above are all with the AlexNet CNN [Krizhevsky et al., 2012]; however, the same MI framework is suitable for even larger CNNs that have produced the most recent top results on ImageNet. Table 4.2 compares the classification accuracy for five different CNN architectures when fine-tuned using my MI framework and the mean aggregation method with 1000×1000 cropped regions. A smaller size was chosen for this experiment to enable each CNN to fit in GPU memory during training. As usual, the whole image was used at test time. A single test of training on half of the data and testing of the other half was used for this experiment.

While it is not possible to distinguish one CNN architecture as better than the others from this limited experiment, it is clear that larger CNNs than AlexNet can successfully be fine-tuned on this data set even with significantly fewer labeled images than ImageNet. This goes to show that a large data set with millions of labeled images is not necessary for fine-tuning a CNN end-to-end; a much smaller set of larger labeled images will suffice when the weak labels are handled in an MI framework.

	Basal vs. non-Basal	ER status	Grade 1 vs. 3
Pre-trained AlexNet SIL-median	0.776	0.772	0.853
Pre-trained AlexNet SIL-quantile	0.799	0.815	0.876
Pre-trained AlexNet MIL-median	0.788	0.807	0.870
Fine-tuned AlexNet	0.831	0.841	0.954
Pre-trained AlexNet SIL-median	0.807	0.823	0.897
Pre-trained VGG16 SIL-quantile	0.824	0.853	0.908
Pre-trained VGG16 MIL-median	0.812	0.846	0.905
Fine-tuned VGG16	0.833	0.879	0.973

Table 4.3: Comparison of classification accuracy with a pre-trained vs. fine-tuned CNN.

4.5.6 Pre-trained vs. Fine-tuned CNN

I also compared classification performance using pre-trained CNN features versus an end-to-end fine-tuned CNN. The pre-trained CNN setup features the SIL-median, SIL-quantile, and MIL-median methods from Chapter 3 and uses the outputs from the fourth convolutional layer of AlexNet [Krizhevsky et al., 2012] or the fourth set of convolutional layers of VGG16 [Simonyan and Zisserman, 2015]. The fine-tuning method used 2000×2000 crops for MI augmentation and mean aggregation. Both were run with half the data for training and half for testing, with the average classification accuracy computed over five random training/test splits. While the fine-tuned CNN method is multi-class, the pre-trained CNN one uses a binary SVM. The multi-class predictions of the CNN are reduced to binary to enable a comparison. However, note that this difference in training setups puts the fine-tuned CNN method at a slight disadvantage. Table 4.3 compares these methods on both AlexNet and VGG16, showing that a fine-tuned CNN consistently outperformed features from a pre-trained one.

Through visualization, I provide insight into how fine-tuning a CNN benefits classification by increasing the discriminative capability of features. I extracted features for each layer of AlexNet and computed the mean value of each feature for each patient (taking the average over the tissue region of each image and further averaging over each core). Figure 4.4 shows t-distributed Stochastic Neighbor Embedding (t-SNE) [Van Der Maaten and Hinton, 2008] plots for the conv4 and dense2 layers of pre-trained and fine-tuned AlexNet. While the pre-trained CNN had some ability to separate Basal from non-Basal, samples of different grades or ER status were not easily separable. The fine-tuned CNN layers did a much better job of distinguishing samples of all classifications, with the separability increased for higher layers of

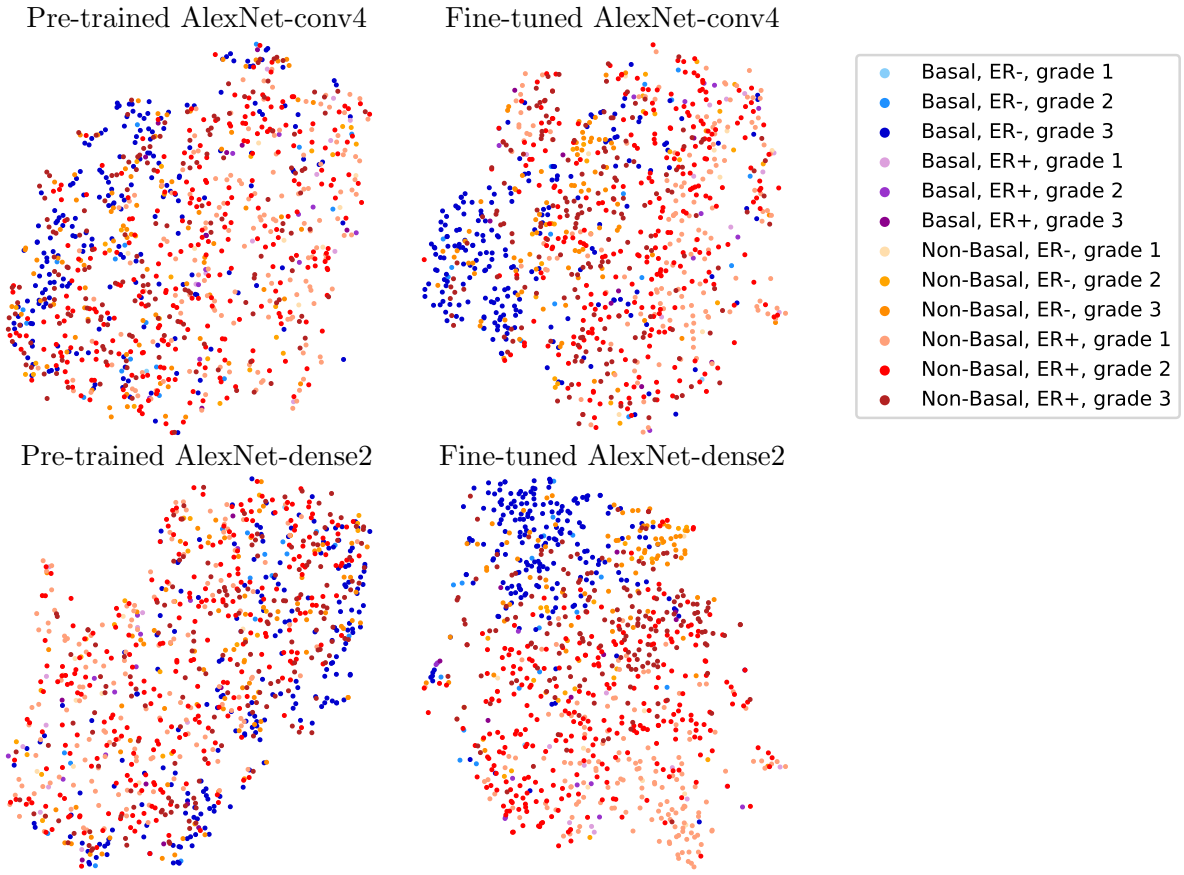

Figure 4.4: t-SNE plots of pre-trained vs. fine-tuned CNN features for the conv4 and dense2 layers of AlexNet. t-SNE is a visualization technique that finds a lower dimensional embedding for high-dimensional data.

the CNN. In particular, note the clusters of Basal samples and non-Basal/ER- samples. The grade of non-Basal/ER+ was also much better separated with the fine-tuned CNN layers. Grade was also the task that received the most benefit from fine-tuning a CNN in Table 4.3. These observations are in line with prior work showing that upper layers of a CNN are increasingly discriminative but only for the application on which they were trained [Yosinski et al., 2014].

4.5.7 Heterogeneity

By computing the class predictions for each instance, I get an idea of each region's contribution to the classification. Figure 4.5 provides a visualization for a sample image where the instance predictions are colored for each class. The $w = 2000$ crop size was used for this example. With the same computation performed over the whole test set, I calculated the proportion

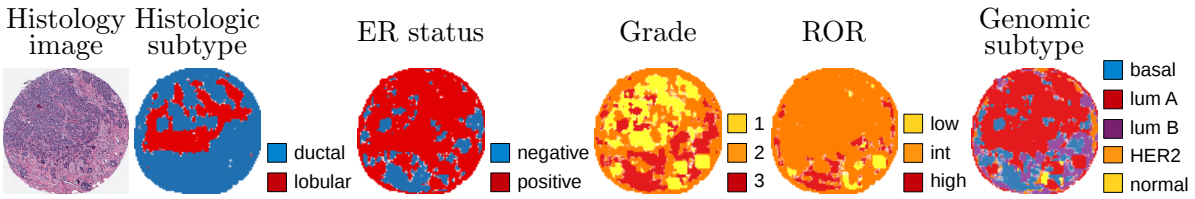

Figure 4.5: Visualization of instance predictions for a sample with ground truth labels of ductal, ER positive, grade 1, low ROR, and luminal A.

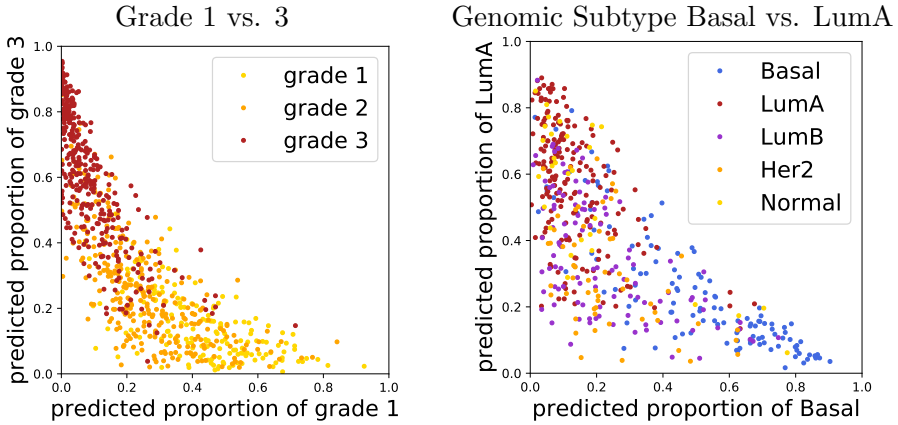

Figure 4.6: Predicted heterogeneity for grade 1 vs. 3 and genomic subtype basal vs. luminal A. The predicted proportion for each class was calculated as the proportion of instances in the sample predicted to be from each class. Test samples for all classes are plotted.

of instances predicted to belong to each class. Figure 4.6 plots the results for grade 1 vs. 3 and genomic subtype basal vs. luminal A. Heterogeneity is expected for grade, as the three tumor grades are not discrete but a continuous spectrum from low to high. On the other hand, the level of heterogeneity to expect for genomic subtype is unknown because no studies have yet assessed genomic subtype from multiple samples within the same tumor. The graph shows a continuous spectrum from basal to luminal A. The luminal B, HER2, and normal samples lie mostly on the luminal A side but with some mixing into the basal side.

4.6 Discussion

I have shown that MI learning while training a CNN is critical in achieving high classification accuracy on large, heterogeneous images. Even with a small number of labeled samples, my model was successful in fine-tuning AlexNet and larger CNNs because of the large size of the images providing plenty of opportunity for MI augmentation. The impact of MI learning indicates that accommodating image heterogeneity is essential. While aggregating instance

predictions with the mean is sufficient for some tasks, quantile aggregation produces a significant improvement for others. This method of fine-tuning a CNN for histology consistently outperforms the methods from Chapter 3, in which pre-trained CNN features are used with an SVM. Instance-level predictions will enable future work studying tumor heterogeneity, perhaps leading to biological insights of tumor progression.

CHAPTER 5: INTEGRATING IMAGE AND GENOMIC FEATURES WITH TASK-DRIVEN DEEP CCA

5.1 Introduction

Thus far, this dissertation has focused on capturing properties of images, but images are not the only data available on tumors. Gene expression is already used to define clinically-relevant subtypes [Parker et al., 2009] and many other forms of genomic, proteomic, and clinical data are available [Ray et al., 2014a]. Image features may also belong to heterogeneous sets as represented by different feature types (Chapter 2) or may be used to characterize different tissue types (e.g., tumor tissue and adjacent stroma [Beck et al., 2011]). Parallel modalities of data are increasingly common in other applications too. They could include other imaging modalities, images and text, audio and video, or parallel texts of different languages. Each modality provides essential information for classification and, when used together, can form a more accurate model. Further, when all modalities are available for training, but only some at test time, the additional training modalities can regularize the model. This is especially import for difficult discriminative tasks such as those with a small training set size or high dimensional input features.

Canonical Correlation Analysis (CCA) is a popular data analysis technique that projects two modalities into a space in which they are maximally correlated [Hotelling, 1936; Bie et al., 2005]. While some of the correlated features extracted by CCA are useful for discriminative tasks, many represent properties that are of no use for classification tasks and obscure correlated information that is beneficial. This problem is magnified with recent non-linear extensions of CCA using deep learning that make significant strides in improving correlation [Andrew et al., 2013; Wang et al., 2015a, 2016b; Chang et al., 2018] but often at the expense of discriminative capability, as will be demonstrated empirically. I present a new set of deep learning techniques to project the data from two modalities to a shared space that is also discriminative.

Variations of CCA have been used for unsupervised data analysis [Andrew et al., 2013; Wang

et al., 2015a; Chandar et al., 2016; Wang et al., 2016b], unsupervised pre-training [Yao et al., 2017; Huang et al., 2018a], content-based retrieval [Li et al., 2003], cross-modal recognition [Kan et al., 2015; Wang et al., 2015a; Chandar et al., 2016; Chang et al., 2018], and multimodal classification [Dorfer et al., 2016b]. Prior work that boosts the discriminative capability of CCA is limited in that it either simply treats the class label as another modality and maximizes the sum correlation over all pairs [Lee et al., 2015; Singanamalli et al., 2014], optimizes discriminative capability for an intermediate representation rather than the final CCA projection [Dorfer et al., 2016b], or is linear only [Lee et al., 2015; Singanamalli et al., 2014; Duan et al., 2016]. I jointly optimize CCA and discriminative objectives by computing the CCA projection within a network layer and applying a task-driven operation such as classification. The sum correlation of the CCA projection and the task-driven objective are optimized as a weighted sum with a hyperparameter for tuning.

Prior work on deep variants of CCA are exclusively focused on large data sets [Andrew et al., 2013; Wang et al., 2015a, 2016b; Chang et al., 2018]. No previous work has studied whether deep CCA can be robust on smaller data sets. On smaller training set sizes, linear CCA overfits the training set and is particularly problematic on high-dimensional low sample size (HDLSS) data in which the feature dimension is greater than the number of samples [Lock et al., 2013]. While other linear methods like Partial Least Squares (PLS) [Wegelin, 2000] are more appropriate in the HDLSS setting, I am not aware of any that have been extended to the non-linear realm with deep learning. PLS maximizes the covariance instead of the correlation, essentially placing orthonormality constraints on the projection weights rather the projections themselves as CCA does. I initially experimented with a form of deep PLS but found the CCA formulation more successful when used in a deep network. By selecting an appropriate hidden layer size, the input feature dimension to CCA can be reduced so that the complications created by an HDLSS input to CCA are diminished. I study my method on small training set sizes and HDLSS data to validate its robustness in these particularly challenging settings, demonstrating that "big data" is not always needed for a successful deep learning solution.

While alternative multimodal approaches to CCA exist, they are typically focused on a reconstruction objective. That is, they transform the input data into a shared space such that the input could be reconstructed - either individually or reconstructing one modality from the

other. Variations of coupled dictionary learning [Shekhar et al., 2014; Xu et al., 2015; Cha et al., 2015; Bahrampour et al., 2015] and auto-encoders [Wang et al., 2015a; Bhatt et al., 2017] have been used, along with further extensions to extract components shared by both modalities and individual components [Ray et al., 2014b; Lock et al., 2013; Zhou et al., 2015; Yang and Michailidis, 2015]. CCA-based objectives, such as the models used in this work, instead learn a transformation to a shared space, without the need for reconstructing the input. This task may be easier and sufficient in producing a representation for cross-modal classification [Wang et al., 2015a].

The contributions of this chapter are in combining the CCA and task-driven objectives in a way that enables end-to-end training. A set of methods is presented, each accomplishing the CCA component in a different way. I demonstrate the effectiveness of these models on a multimodal variation of MNIST adapted for study across training set size and input feature dimension, showing that task-driven deep CCA significantly improves cross-modal classification accuracy and is more robust to small training set size and HDLSS data than alternative methods. I also provide validation on two cancer imaging and genomic data sets, the Carolina Breast Cancer Study (CBCS) and The Cancer Genome Atlas (TCGA); the first has a small training set size, and the second is HDLSS data. Once again, improvements in cross-modal classification are demonstrated. Further experiments show that these CCA-based models can also be used as a means to regularize a model when two modalities are available for training but only one is available at test time. This technique is shown on the CBCS data set to improve classification accuracy from images.

Section 5.2 presents background material on CCA and Deep CCA. My proposed methods for task-driven deep CCA are detailed in Section 5.3. Section 5.4 discusses related work in more detail. Finally, experimental validation is presented in Section 5.5.

5.2 Background: CCA and Deep CCA

CCA and its non-linear variants are unsupervised methods that find the shared signal between a pair of modalities by minimizing the difference between orthonormal projections of the two modalities. Let $X_1 \epsilon \mathbb{R}^{d_1 \times n}$ and $X_2 \epsilon \mathbb{R}^{d_2 \times n}$ be mean centered input data matrices from two

different modalities with n samples and d_1 or d_2 features.

CCA. CCA maximizes the correlation between linear projections $a_1 = w_1^T X_1$ and $a_2 = w_2^T X_2$, where w_1 and w_2 are projection vectors [Hotelling, 1936]. The first canonical direction is found by maximizing the correlation

$$\underset{w_1,w_2}{\operatorname{argmax}} \operatorname{corr}(w_1^T X_1, w_2^T X_2).$$

Subsequent projections are found by maximizing the same correlation but in orthogonal directions: $(w_1^{(i)})^T w_1^{(j)} = (w_2^{(i)})^T w_2^{(j)} = 0$ for all $i \neq j$.

The projection vectors $W_1 = [w_1^{(1)}, ..., w_1^{(k)}]$ and $W_2 = [w_2^{(1)}, ..., w_2^{(k)}]$ ($k \leq \min(d_1, d_2)$) can be found by first reformulating as the trace with orthonormality constraints:

$$\underset{W_1,W_2}{\operatorname{argmax}} \operatorname{tr}(W_1^T \Sigma_{12} W_2) \quad \text{s.t. } W_1^T \Sigma_1 W_1 = W_2^T \Sigma_2 W_2 = I$$

for covariance matrices $\Sigma_1 = \frac{1}{n-1} X_1 X_1^T$ and $\Sigma_2 = \frac{1}{n-1} X_2 X_2^T$ and cross-covariance matrix $\Sigma_{12} = \frac{1}{n-1} X_1 X_2^T$. Let $T = \Sigma_1^{-1/2} \Sigma_{12} \Sigma_2^{-1/2}$ and SVD $T = U_1 \operatorname{diag}(\sigma) U_2^T$ with singular values $\sigma = [\sigma_1, ..., \sigma_k]$ in descending order. W_1 and W_2 are computed from the top k left and right singular vectors of T:

$$W_1 = \Sigma_1^{-1/2} U_1^{(1:k)} \qquad W_2 = \Sigma_2^{-1/2} U_2^{(1:k)} \tag{5.1}$$

where $U^{(1:k)}$ is the k first columns of matrix U. The sum correlation in the projection space is equivalent to the sum of the top k singular values:

$$\sum_{i=1}^{k} \operatorname{corr}((w_1^{(i)})^T X_1, (w_2^{(i)})^T X_2) = \sum_{i=1}^{k} \sigma_i^2 \ . \tag{5.2}$$

A regularization parameter r can be used to ensure that the covariance matrices are positive definite and to increase stability for HDLSS data ($d_1, d_2 >> n$). In this Regularized CCA (RCCA) method, the covariance matrices are computed as $\hat{\Sigma}_1 = \frac{1}{n-1} X_1 X_1^T + rI$ and $\hat{\Sigma}_2 = \frac{1}{n-1} X_2 X_2^T + rI$ for regularization parameter r and identity matrix I [Bilenko and Gallant, 2016].

DCCA. Deep CCA replaces the linear projections with feed-forward networks f_1 and f_2, using parameters θ_1 and θ_2 and producing activations $A_1 = f_1(X_1; \theta_1)$ and $A_2 = f_2(X_2; \theta_2)$, respectively (assumed to be mean centered) [Andrew et al., 2013]. Matrices $A_1, A_2 \epsilon \mathbb{R}^{d_o \times n}$ are the output activations from the networks f_1 and f_2 with d_o features on the output layer. Figure 5.1a shows the network structure.

DCCA optimizes a trace objective with some constraints:

$$\underset{W_1, W_2, \theta_1, \theta_2}{\operatorname{argmax}} \operatorname{tr}(W_1^T \Sigma_{12} W_2^T) \quad \text{s.t. } W_1^T \Sigma_1 W_1 = W_2^T \Sigma_2 W_2 = I \tag{5.3}$$

where $\Sigma_1 = \frac{1}{n-1} A_1 A_1^T + rI$, $\Sigma_2 = \frac{1}{n-1} A_2 A_2^T + rI$, and $\Sigma_{12} = \frac{1}{n-1} A_1 A_2^T$ are now computed using network outputs A_1 and A_2 and where the regularization parameter r is used in the same way as RCCA. The solution for W_1 and W_2 can be computed using SVD just as with linear CCA (Equation 5.1). When $k = d_o$ (the number of CCA components is equal to the number of features in A_1 and A_2), optimizing the sum correlation in the projection space (Equation 5.2) is equivalent to optimizing the matrix trace norm $||T||_{\text{tr}} = \operatorname{tr}(T^T T)^{1/2}$ where $T = \Sigma_1^{-1/2} \Sigma_{12} \Sigma_2^{-1/2}$ as for CCA. DCCA optimizes this trace norm objective (TNO) directly, without a need to compute the CCA projection within the network. The TNO is optimized first, followed by a linear CCA operation before downstream tasks like classification are performed.

CCA and DCCA are both unsupervised methods to learn a projection to a space where the data is maximally correlated. This representation can then be used for other tasks like training a classifier.

5.3 Task-driven Deep CCA

CCA finds a projection in which the modalities are maximally correlated; however, only some of these directions are useful for discriminative purposes. I propose a set of methods to add a task-driven term to the optimization objective in order to find a projection that is also beneficial for classification. While pieces of these methods have been presented previously, their integration into a task-driven deep CCA model is unique in this work.

Deep Neural Networks (DNNs) $A_1 = f_1(X_1; \theta_1)$ and $A_2 = f_2(X_2; \theta_2)$ are applied to each modality X_1 and X_2, and the task-driven objective operates on the outputs A_1 and A_2, respec-

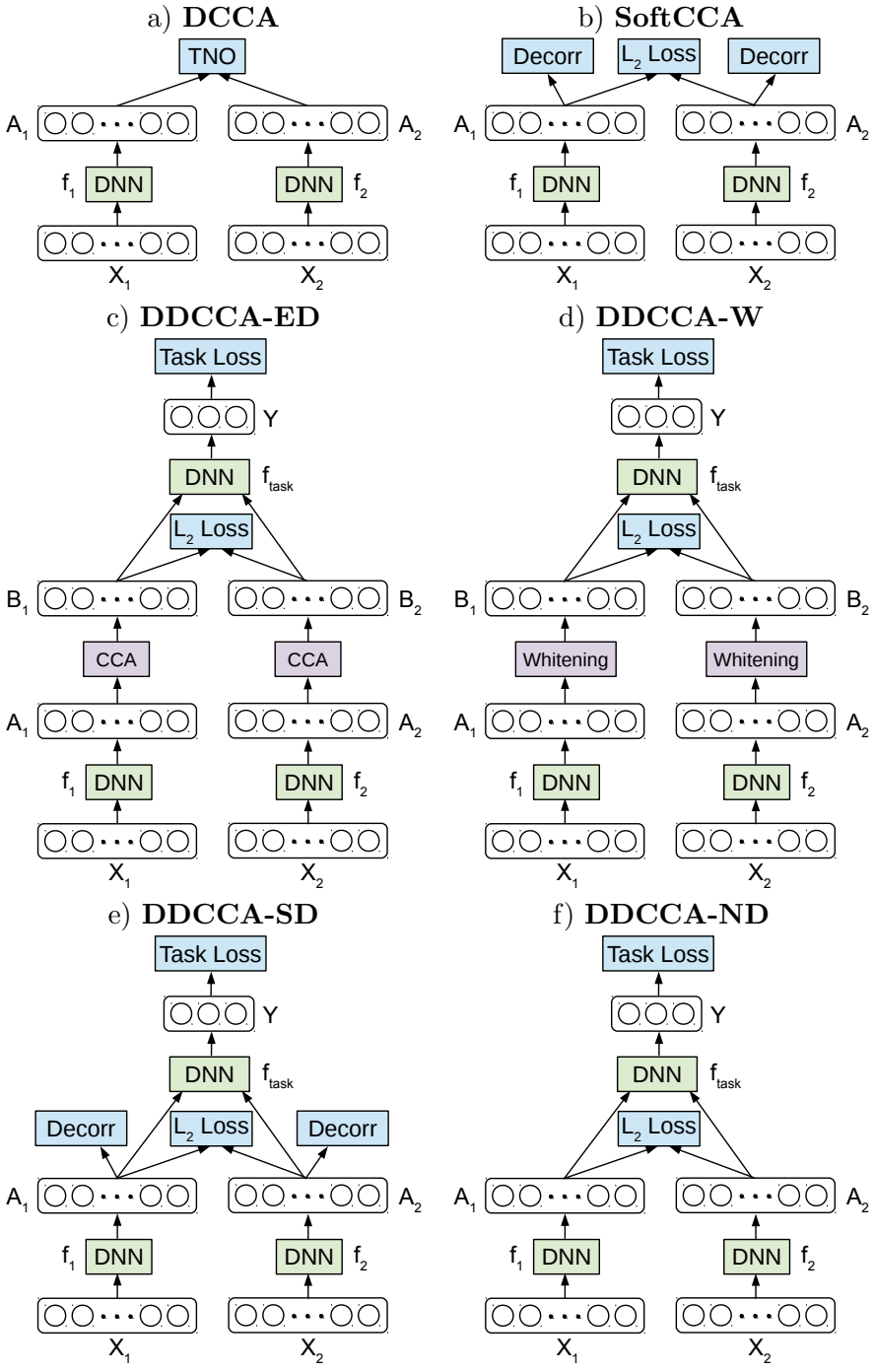

Figure 5.1: Deep CCA architectures used in this work: a) unsupervised DCCA using the trace norm objective (TNO) [Andrew et al., 2013], b) unsupervised SoftCCA using soft decorrelation [Chang et al., 2018], c) DDCCA-ED using eigendecomposition, d) DDCCA-W using whitening, e) DDCCA-SD using soft decorrelation, f) DDCCA-ND with no explicit decorrelation.

tively. Task-driven functions $f_{\text{task1}}(A_1; \theta_{task})$ and $f_{\text{task2}}(A_2; \theta_{\text{task}})$, such as the softmax, then perform the discriminative task using the activations A_1 and A_2. A_1 and A_2 are optimized to be maximally correlated, so the parameters θ_{task} are shared between the two functions.

While DCCA provides a means to optimize the sum correlation through the TNO, the CCA projection itself is computed only after optimization is complete. In developing a task-driven form of deep CCA that discriminates based on the CCA projection and is trained end-to-end, we must compute this projection within the network. Four approaches will be considered in this section. Each integrates the CCA projection into the DNN in a different way, while still maximizing the sum correlation between modalities. The trace objective for DCCA in Equation 5.3 can be rewritten as minimizing the ℓ_2 distance between the projections

$$\underset{W_1, W_2, \theta_1, \theta_2}{\text{argmax}} \; ||W_1^T A_1 - W_2^T A_2||_F^2 \tag{5.4}$$

as long as $W_1^T A_1$ and $W_2^T A_2$ are normalized to one:

$$(w_1^{(i)})^T A_1 A_1^T w_1^{(i)} = 1 \quad \text{and} \quad (w_2^{(i)})^T A_2 A_2^T w_2^{(i)} = 1 \text{ for all } i$$

where $w_1^{(i)}$ and $w_2^{(i)}$ are the i-th columns of W_1 and W_2, respectively [Li et al., 2003]. This new objective

$$\mathcal{L}_{\ell_2 \text{ dist}}(A_1, A_2) = ||W_1^T A_1 - W_2^T A_2||_F^2 \tag{5.5}$$

is now separable across mini-batches and will be used by each method proposed below.

Transforming the CCA objective into one that is separable across mini-batches is an important first step. Each method that I present maintains this property. While large mini-batch implementations of stochastic gradient optimization increase the opportunity for parallelism and therefore increased computational efficiency, small batch training increases the range of suitable learning rates and improves test accuracy [Masters and Luschi, 2018]. The orthogonality constraints of CCA are applied differently in each method that I develop, and each is formulated such that it can be run on any batch size.

I present four different formulations for Discriminative Deep CCA (DDCCA):

1) **DDCCA-ED** computes the CCA projection within the network using eigendecomposi-

tion,

2) **DDCCA-W** uses whitening to achieve orthogonality within each modality,

3) **DDCCA-SD** relaxes the orthogonality constraints with regularization or soft decorrelation, and

4) **DDCCA-ND** does not make any specific effort to decorrelate, relying upon the task-driven objective to encourage this.

Each maximizes the sum correlation using the ℓ_2 distance objective in Equation 5.4 but accomplishes decorrelation in a different manner. Figure 5.1 shows the network architecture for each of these models.

These methods are robust to small training set sizes and HDLSS data because the DNN downsizes the data dimensionality before applying CCA, reducing the effects of overfitting seen with linear CCA. Further, the task-driven component provides additional information that regularizes the model.

1) DDCCA-ED: CCA Projection Layer by Eigendecomposition. The CCA projection is typically computed using SVD, requiring a gradient computation for optimization with backpropagation; however, a gradient for SVD is not provided in auto-differentiating math compilers like Theano. Following Dorfer et al., I use eigendecomposition to compute the CCA projection for each modality [Dorfer et al., 2018], enabling the use of backpropagation for end-to-end training. Further, I use a stochastic approximation of the covariance and cross-covariance matrices to enable computation on smaller batch sizes.

Given activations A_1 and A_2 from two feed-forward networks, as defined previously, the CCA projection can be computed with eigendecomposition. Let the covariance and cross-covariance matrices be $\Sigma_1 = \frac{1}{n-1}A_1A_1^T + rI$, $\Sigma_2 = \frac{1}{n-1}A_2A_2^T + rI$, and $\Sigma_{12} = \frac{1}{n-1}A_1A_2^T$, respectively. Recall from Equation 5.1 that the CCA projections are computed from the SVD $T = \Sigma_1^{-1/2}\Sigma_{12}\Sigma_2^{1/2} = U\text{diag}(d)V^T$. Following Dorfer et al., I use two symmetric eigendecompositions instead of SVD: $TT^T = U_1\text{diag}(e_1)U_1^T$ and $T^TT = U_2\text{diag}(e_2)U_2^T$ where e_1 and e_2 contain the eigenvalues and U_1 and U_2 contain the respective eigenvectors [Dorfer et al., 2018]. The CCA projections can now be computed as $W_1 = \Sigma_1^{-1/2}U_1$ and $W_2 = \Sigma_2^{-1/2}U_2$. This method is equivalent to the SVD but

with the possibility of a flipped sign that is easily resolved by flipping signs of W_1 to match W_2 with positive correlations. While Dorfer et al. computed $\Sigma_1^{-1/2}$ and $\Sigma_2^{-1/2}$ with the Cholesky factorization, I found this to be numerically unstable and instead use eigendecomposition.

These operations are clearly dependent upon covariance matrices Σ_1 and Σ_2; however, computation of the covariance matrix is not separable across batches. Following Wang et al. [Wang et al., 2016b] and Chang et al. [Chang et al., 2018], I use a stochastic approximation updated at each batch k as

$$\Sigma^{(k)} = \alpha\Sigma^{(k-1)} + (1-\alpha)\Sigma^b \tag{5.6}$$

where $\Sigma^{(0)} = I$, Σ^b is the covariance matrix calculated for the current batch, and α is the momentum hyperparameter. The CCA projection is then calculated from $\Sigma^{(k)}$ for the current batch:

$$B_1 = f_{CCA1}(A_1) = (\Sigma_1^{(k)})^{-1/2}U_1^{(k)}A_1 \qquad B_2 = f_{CCA2}(A_2) = (\Sigma_2^{(k)})^{-1/2}U_2^{(k)}A_2$$

where $\Sigma_1^{(k)}$, $\Sigma_2^{(k)}$, $U_1^{(k)}$, and $U_2^{(k)}$ are computed as above for the current batch.

The loss function for DDCCA-ED integrates both the correlation and task-driven objectives:

$$\mathcal{L}_{\text{task}}(B_1, Y) + \mathcal{L}_{\text{task}}(B_2, Y) + \lambda\mathcal{L}_{\ell_2 \text{ dist}}(B_1, B_2)$$

with hyperparameter λ to adjust the weight of the last term. At test time, matrices $U_1^{(k)}$ and $U_2^{(k)}$ from the last training batch are used. This method is detailed more explicitly in Algorithm 1.

2) DDCCA-W: Whitening. Deep CCA has two components: maximizing the sum correlation (or, equivalently, minimizing the ℓ_2 distance) and enforcing orthonormality constraints. DDCCA-W takes a different approach to applying the orthonormality constraints by using whitening to decorrelate A_1 and A_2. I use Decorrelated Batch Normalization (DBN) [Huang et al., 2018b] as inspiration for the model I present here. DBN regularizes a deep model by decorrelating features with whitening, that is, applying a transformation $B = UA$ to make B orthonormal: $BB^T = I$. The covariance matrix of A is computed as $\Sigma = \frac{1}{n-1}AA^T$. Any matrix

Algorithm 1 Forward pass of CCA projection layer using eigendecomposition.

Input: $A_1, A_2 \epsilon \mathbb{R}^{d_o \times n}$

Hyperparameters: mini-batch size m, momentum α

Parameters of layer: μ_1, μ_2, Σ_1, Σ_2, Σ_{12}

if training **then**

$\mu_1 \leftarrow \alpha\mu_1 + (1-\alpha)\frac{1}{m}A_1\ 1_{n\times 1}$ ▷ Update running mean with batch mean

$\mu_2 \leftarrow \alpha\mu_2 + (1-\alpha)\frac{1}{m}A_2\ 1_{n\times 1}$

$\bar{A}_1 = A_1 - \mu_1$ ▷ Mean center data

$\bar{A}_2 = A_2 - \mu_2$

$\Sigma_1 \leftarrow \alpha\Sigma_1 + (1-\alpha)\frac{1}{m-1}\bar{A}_1\bar{A}_1^T$ ▷ Update running cov matrix with batch cov

$\Sigma_2 \leftarrow \alpha\Sigma_2 + (1-\alpha)\frac{1}{m-1}\bar{A}_2\bar{A}_2^T$

$\Sigma_{12} \leftarrow \alpha\Sigma_{12} + (1-\alpha)\frac{1}{m-1}\bar{A}_1\bar{A}_2^T$

$\hat{\Sigma}_1 \leftarrow \Sigma_1 + rI$ ▷ Add regularization

$\hat{\Sigma}_2 \leftarrow \Sigma_2 + rI$

$\Lambda_1, V_1 \leftarrow \text{eig}(\hat{\Sigma}_1)$ ▷ Compute eigendecomposition

$\Lambda_2, V_2 \leftarrow \text{eig}(\hat{\Sigma}_2)$

$\Sigma_1^{-1/2} \leftarrow V_1\Lambda_1^{-1/2}V_1^T$ ▷ Compute matrix square root inverse

$\Sigma_2^{-1/2} \leftarrow V_2\Lambda_2^{-1/2}V_2^T$

$T \leftarrow \Sigma_1^{-1/2}\Sigma_{12}\Sigma_2^{-1/2}$ ▷ Compute matrix T

$\Lambda, U_1 \leftarrow \text{eig}(TT^T)$ ▷ Compute eigenvectors of TT^T

$\Lambda, U_2 \leftarrow \text{eig}(T^TT)$ ▷ Compute eigenvectors of T^TT

$W_1 \leftarrow \Sigma_1^{-1/2}U_1$ ▷ Compute projection matrices

$W_2 \leftarrow \Sigma_2^{-1/2}U_2$

$W_1 \leftarrow W_1\text{sgn}(\text{diag}(W_1^T\Sigma_{12}W_2))$ ▷ Flip sign when necessary

else

$\bar{A}_1 \leftarrow A_1 - \mu_1$ ▷ Mean center data

$\bar{A}_2 \leftarrow A_2 - \mu_2$

end if

$B_1 \leftarrow W_1\bar{A}_1$ ▷ Compute CCA projection using parameters estimated during training

$B_2 \leftarrow W_2\bar{A}_2$

return B_1, B_2

$U \epsilon \mathbb{R}^{d_o \times d_o}$ that satisfies the condition $U^T U = \Sigma^{-1}$ whitens the data; however, U is only defined up to a rotation, so it is not unique. PCA whitening uses the eigendecomposition of covariance matrix Σ: $U_{PCA} = \Lambda^{-1/2} V^T$ for $\Lambda = \text{diag}(\lambda_1, ..., \lambda_{d_o})$ and $V = [v_1, ..., v_1]$ where (λ_i, v_i) are the eigenvalue, eigenvector pairs of Σ. However, PCA whitening suffers from stochastic axis swapping in which the neurons are not stable from one batch to the next [Huang et al., 2018b]. I instead use Zero-phase Component Analysis (ZCA) whitening as recommended by Huang et al. and used in their DBN method [Huang et al., 2018b]. ZCA uses the transformation $U_{ZCA} = V \Lambda^{-1/2} V^T$ in which PCA whitening is first applied, followed by a rotation back to the original space. Adding the rotation V brings the whitened data B as close as possible to the original data A [Kessy et al., 2015]. Figure 2.3 in Chapter 2 provides a visual overview of the operations in ZCA whitening.

Computation of U_{ZCA} is clearly dependent upon covariance matrix Σ. While Huang et al. use a running average of transformation matrix U_{ZCA} over mini-batches [Huang et al., 2018b], I apply this stochastic approximation to covariance matrix Σ for each modality using the update for batch k in Equation 5.6. I then compute the ZCA transformation from $\Sigma^{(k)}$ to do whitening:

$$B = f_{\text{ZCA}}(A) = U_{\text{ZCA}}^{(k)} A \ .$$

At test time, matrix $U^{(k)}$ from the last training batch is used.

Algorithm 2 describes the forward pass for ZCA whitening in more detail. Through this stochastic method of decorrelation using whitening, CCA can be applied within a DNN and combined with a discriminative term. The loss function for DDCCA-W integrates both the correlation and task-driven objectives with decorrelation performed by whitening:

$$\mathcal{L}_{\text{task}}(B_1, Y) + \mathcal{L}_{\text{task}}(B_2, Y) + \lambda \mathcal{L}_{\ell_2 \text{ dist}}(B_1, B_2) \ .$$

The whole network is trained end-to-end.

3) DDCCA-SD: Soft Decorrelation. While fully independent components may be beneficial in regularizing a DNN on some data sets, a softer decorrelation may be more suitable on others. Thus, I loosen the orthogonality constraints by applying them as regularization. DeCov

Algorithm 2 Forward pass of whitening layer.

Input: $A \epsilon \mathbb{R}^{d_o \times n}$
Hyperparameters: mini-batch size m, momentum α
Parameters of layer: μ, Σ

if training **then**
  $\mu \leftarrow \alpha\mu + (1-\alpha)\frac{1}{m} A\ 1_{n\times 1}$ ▷ Update running mean with batch mean
  $\bar{A} = A - \mu$ ▷ Mean center data
  $\Sigma \leftarrow \alpha\Sigma + (1-\alpha)\frac{1}{m-1}\bar{A}_1\bar{A}_2^T$ ▷ Update running cov matrix with batch cov
  $\hat{\Sigma} \leftarrow \Sigma + \epsilon I$ ▷ Add ϵI for numerical stability
  $\Lambda, V \leftarrow \text{eig}(\hat{\Sigma})$ ▷ Compute eigendecomposition
  $W \leftarrow V\Sigma^{-1/2}V^T$ ▷ Compute transformation for ZCA whitening
else
  $\bar{A} \leftarrow A - \mu$ ▷ Mean center data
end if
$B \leftarrow W\bar{A}$ ▷ Compute ZCA whitened output using parameters estimated during training
return B

regularization performs such a decorrelation but operates on batches independently [Cogswell et al., 2016]. Stochastic Decorrelation Loss (SDL) instead uses a running average to update the covariance matrix over batches and is the formulation that I use here [Chang et al., 2018]. Chang et al. also use SDL in a CCA-type model; however, they have no task-driven component.

I selected an ℓ_1 penalty and formulate the decorrelation loss using each modality's covariance matrix Σ. I approximate Σ across batches with Equation 5.6 and express the decorrelation loss for batch k as

$$\mathcal{L}_{\text{Decorr}}(\Sigma) = \sum_{i=1}^{d_o}\sum_{j\neq i}^{d_o} |\Sigma_{i,j}| \ .$$

The loss function for this formulation is then

$$\mathcal{L}_{\text{task}}(A_1, Y) + \mathcal{L}_{\text{task}}(A_2, Y) + \lambda_1 \mathcal{L}_{\ell_2 \text{ dist}}(A_1, A_2) + \lambda_2 \left(\mathcal{L}_{\text{Decorr}}(\Sigma_1^{(k)}) + \mathcal{L}_{\text{Decorr}}(\Sigma_2^{(k)})\right) \ .$$

4) DDCCA-ND: No Decorrelation. When CCA is used in an unsupervised manner, some sort of orthogonality constraints or decorrelation is necessary to ensure that networks f_1 and f_2 do not simply produce multiple copies of the same feature. While this result could maximize the sum correlation, it is not helpful in capturing useful projections. In the task-driven setting, the SDL term helps to ensure that the features in f_1 and f_2 capture properties that are discriminative and therefore not replicates of the same information. In this formulation I remove

the decorrelation term from DDCCA-SD entirely, forming an even simpler objective:

$$\mathcal{L}_{\text{task}}(A_1, Y) + \mathcal{L}_{\text{task}}(A_2, Y) + \lambda \mathcal{L}_{\ell_2 \text{ dist}}(A_1, A_2) \ .$$

This allows me to test whether the soft decorrelation term provides a beneficial regularization in a task-driven model.

Summary. The proposed task-driven CCA methods optimize two objectives: maximizing the sum correlation between modalities in the projected space (by minimizing the ℓ_2 distance) and a task-driven objective. This model is flexible, in that the task-driven goal can be for classification [Krizhevsky et al., 2012; Dorfer et al., 2016a], regression [Katzman et al., 2016], clustering [Caron et al., 2018], or some other task entirely. Both objectives are optimized simultaneously with end-to-end training.

Combining the task-driven and CCA objectives in a deep model is the main contribution of this chapter. Prior CCA-based methods do not use label information [Andrew et al., 2013; Chang et al., 2018], do not compute the final CCA projection within the network [Andrew et al., 2013; Dorfer et al., 2016b], or are linear only [Singanamalli et al., 2014; Lee et al., 2015; Kan et al., 2015; Duan et al., 2016]. Each of the four models further builds upon prior work in the following ways:

- **DDCCA-ED** - This model computes the CCA projection within the network, following the method by Dorfer et al. [Dorfer et al., 2018]. I further adapt their CCA projection algorithm by computing the covariance and cross-covariance matrices as a running average over mini-batches to enable small batch training.

- **DDCCA-W** - Instead of explicitly computing the CCA projection, this model relies upon ℓ_2 distance minimization combined with whitening to achieve the orthogonality constraints of CCA. ZCA whitening has previously been used in a DNN as Decorrelated Batch Normalization [Huang et al., 2018b]; however, instead of applying a running average computation over mini-batches to the transformation matrix, I apply it to the covariance matrices.

- **DDCCA-SD** - While the previous two models orthogonalize the projections from each modality, this method loosens this constraint by applying it as regularization. This technique has been done before in the form of SoftCCA which uses a Soft Decorrelation Loss by computing the covariance matrices as a running average over mini-batches [Chang et al., 2018]. My implementation is very much the same as SoftCCA but with the addition of the task-driven objective.

- **DDCCA-ND** - In this model, I remove all methods to explicitly decorrelate each modality. This is simply a special case of DDCCA-SD, with the weight on the decorrelation term set to zero. The task-driven objective still encourages some amount of decorrelation in order to achieve its goal.

All methods are designed to be run on any batch size, enabling the optimization benefits of improved test performance and stable convergence from small batch training [Masters and Luschi, 2018].

5.4 Related Work

CCA on High-Dimensional Data. CCA finds the shared signal between a pair of modalities by maximizing the correlation of a set of orthogonal projections [Hotelling, 1936]. When the feature dimension is much larger than the number of samples, CCA becomes unstable. Regularized CCA provides one solution by adding a regularizing parameter to the covariance matrices [Bilenko and Gallant, 2016]. When the covariance matrices are maximally regularized, the objective shifts to maximizing the covariance of the modality projections instead of the correlation - known as Partial Least Squares (PLS) [Wegelin, 2000] or Cross-modal Factor Analysis (CFA) [Li et al., 2003]. CCA has been extended to the non-linear case with Kernel CCA [Lai and Fyfe, 2000] and Deep CCA (DCCA) [Andrew et al., 2013]. Using DCCA with appropriately chosen hidden layer dimensions, high-dimensional input features can be reduced, providing an alternative solution to the high-dimensional problem. I study deep variations of CCA on HDLSS data, for which I am unaware of any prior work.

CCA Approximation with Deep Learning. Recent work on deep learning has shown that smaller batches increase stability and produce better results than larger ones due to the more up-to-date gradient calculation [Masters and Luschi, 2018]. DCCA was developed for very large batches because correlation is not separable across batches, whereas I present a stochastic approximation that can be run on any batch size and will be shown to perform better with a smaller batch size than a larger one. Wang et al. also tackle this problem and approximate the covariance matrices across batches; however, while they perform whitening to approximate the orthogonality constraint, it is not clear what type of whitening is used [Wang et al., 2016b]. Further, DCCA does not explicitly compute the final modality projections within the network but instead optimizes an equivalent loss function [Andrew et al., 2013]. This means that the projections cannot be used to optimize another task simultaneously with end-to-end training. Dorfer et al. compute this projection in the network, but they only optimize a task-specific objective, not the sum correlation [Dorfer et al., 2018]. My deep CCA formulations compute the projection within the network and simultaneously optimize a task-driven objective.

CCA Relaxation. CCA constrains its components to be orthogonal. Chander et al. remove this orthogonality completely in their Correlational Neural Network model but add cross-modal autoencoder terms [Chandar et al., 2016]. As a middle ground, Soft CCA relaxes the orthogonality constraints by decorrelating features with regularization [Chang et al., 2018] in a similar manner to DeCov regularization on unimodal data [Cogswell et al., 2016]. I present DDCCA approaches for both orthogonal projections and soft decorrelation.

Supervised CCA. Although CCA has shown utility for discriminative tasks, this particular decomposition is not necessarily optimal for classification purposes because it can also extract features that are correlated but not discriminative. My experiments will make it clear that maximizing the correlation objective too much can degrade discriminative tasks. CCA has previously been extended to the supervised case by maximizing the total correlation between each modality and the training labels in addition to each pair of modalities [Lee et al., 2015; Singanamalli et al., 2014] and by maximizing the separation of classes [Kan et al., 2015; Dorfer et al., 2016b]. Although these methods incorporate the class labels, they do not directly optimize

for the classification task with end-to-end training. Duan at al. jointly optimize a CCA-type model with an SVM for classification; however, their model only supports linear projections [Duan et al., 2016]. My DDCCA method simultaneously optimizes CCA and a task-driven objective in a deep network to capture non-linear representations. The CCA objective enables cross-modal classification in which a model is trained on one modality and the other modality is used at test time. Further, analysis on the data in the shared space can provide new insights, such as with the visualizations I present in Figures 5.5 and 5.6.

5.5 Experiments

The experiments that follow study the performance of my proposed methods and compare with other related methods, including the following:

- **CCA** - Canonical Correlation Analysis [Hotelling, 1936; Bie et al., 2005; Wegelin, 2000]
- **RCCA** - Regularized Canonical Correlation Analysis which adds a regularizing parameter to the covariance matrix in CCA [Bilenko and Gallant, 2016]
- **PLS-SVD** - Partial Least Squares using SVD [Wegelin, 2000]
- **DCCA** - Deep CCA that extends CCA with deep learning [Andrew et al., 2013], modified to use ReLu activation and batch normalization
- **SoftCCA** - Deep CCA with soft decorrelation [Chang et al., 2018]

In order to test the cross-modal classification accuracy, I used each model to compute the projection for each modality. I then trained a linear SVM on one modality projection and used the other modality projection at test time. While the task-driven methods presented in this work learn a classifier within the model, this test setup enables a comparison with the unsupervised forms of CCA. Experimental results showed that the DDCCA methods proposed in this chapter outperformed the alternatives listed above and are more robust to small training set sizes and HDLSS data.

5.5.1 Implementation Details

Each layer of my network consists of a fully connected layer, followed by the ReLu activation function and batch normalization [Ioffe and Szegedy, 2015]. I used the Nadam optimizer and trained for 200 epochs on MNIST and 400 on the other data sets. Hidden layer size was set to 500 for MNIST and 200 for CBCS and TCGA-BRCA, with 0 to 4 layers for each modality. I learned the hyperparameters on a validation set by random search, including number of hidden layers, loss function weight λ, momentum α, ℓ_2 regularizer, learning rate, and batch size. I used Keras with the Theano backend.

5.5.2 Synthetic Examples with MNIST Split

I formed a multimodal data set from the MNIST handwritten image data set [LeCun, 1998]. Following Andrew et al., I split each 28×28 image in half horizontally, creating left and right modalities that are each 14×28 pixels [Andrew et al., 2013]. Each modality was flattened into a vector with 392 features. The full data set consists of 60k training images and 10k test images. I used a random set of up to 50k for training and the remaining training images for validation. I used the full 10k image test set. I trained a DNN with 4 hidden layers of size 500 and an output layer of size 50. I report the mean sum correlation or classification accuracy over five randomly selected training/validation sets; the test set always remained the same.

Correlation vs. Classification Accuracy. I first demonstrate the importance of adding a task-driven component to deep CCA by showing that maximizing the sum correlation between modalities is not sufficient and can even be misleading. Figure 5.2 plots the sum correlation vs. cross-modal classification accuracy across many different hyperparameter settings for DCCA [Andrew et al., 2013] and SoftCCA [Chang et al., 2018]. I used 50 components for each and thus the maximum possible sum correlation was 50. The sum correlation was measured after applying linear CCA to the network projections to ensure that components are independent.

With DCCA, a larger correlation tended to produce a larger classification accuracy, but there was still a large variance in classification accuracy amongst hyperparameter settings that produced a similar sum correlation. Take, for example, the two farthest right points in the

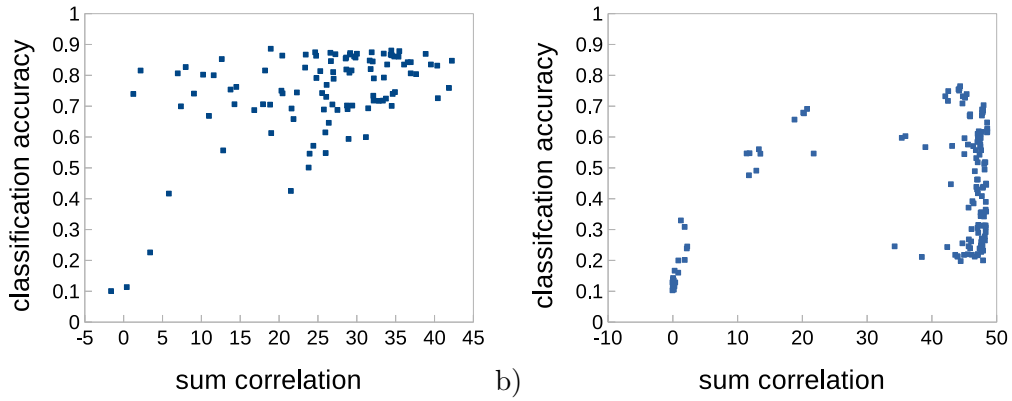

Figure 5.2: Sum correlation vs. cross-modal classification accuracy across many different hyper-parameter settings on a training set size of 10,000. a) DCCA [Andrew et al., 2013], b) SoftCCA [Chang et al., 2018]

plot: their classification accuracy differs by 10%, and they are not even the points with the best classification accuracy. Further, some settings with a very small sum correlation resulted in a very high classification accuracy. The pattern is rather different for SoftCCA. There was an increase in classification accuracy as sum correlation increased but only up to a point. For higher sum correlations, the classification accuracy varied even more from 20% to 80%. Optimizing for sum correlation alone does not guarantee a model with the highest cross-modal classification accuracy.

The Importance of Small Batch Size. Each of the methods developed in this chapter was designed to be run on any batch size in order get the best cross-modal classification accuracy from small batches. This experiment verifies that small batch training is best for all four task-driven deep CCA methods. Figure 5.3 plots the batch size vs. classification accuracy for a training set size of 10,000. Batch sizes of 10,000, 1000, and 100 were tested, showing that a batch size of 100 produced a small improvement for most of the methods and a large improvement for DDCCA-ED. The latter method is particularly susceptible to poor sub-optimal convergence with a large batch size. Even smaller batch sizes could be tested, although they result in an increased runtime. Masters and Luschi found the best performance was obtained for a batch size between 2 and 32 in their experiments [Masters and Luschi, 2018].

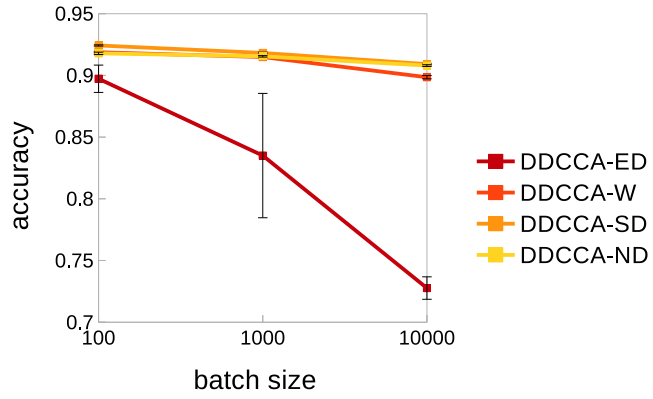

Figure 5.3: The effect of batch size on classification accuracy for each task-driven multimodal method on MNIST split with a training set size of 10,000.

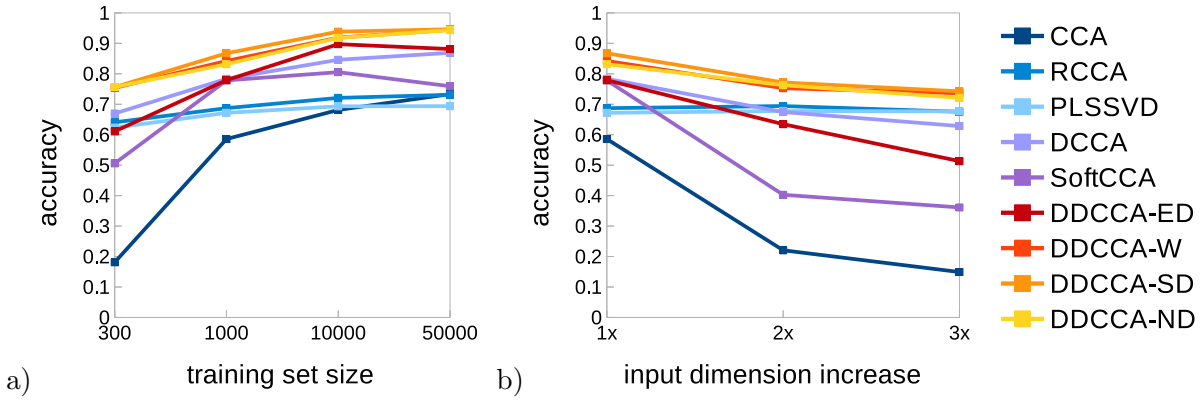

Figure 5.4: The effect of training set size and input dimension size on classification accuracy for each multimodal method on MNIST split. a) The training set size was manipulated by selecting a random subset of examples for training. b) With a training set size of 10,000, the input dimension was increased by a given factor by adding extra background surrounding the digit.

Training Set Size. I manipulated the training set size in order to study the robustness of my methods to smaller data sets. Figure 5.4a shows the cross-modal classification accuracy for training set sizes of $n = 300$, 1000, 10,000, and 50,000. While it is expected that performance will decrease for smaller training set sizes, some methods are more susceptible to this degradation than others. As expected, the classification accuracy with CCA dropped significantly for $n = 300$ and 1000 as the covariance and cross-covariance matrices were not stable and the training data was overfit. SoftCCA was also particularly susceptible for $n = 300$. Prior work on this method did not test such small training set sizes [Chang et al., 2018].

Across all training set sizes, DDCCA-W and DDCCA-SD were consistently the top performers, for example, increasing classification accuracy from 78.3% to 86.7% for $n = 1000$.

Increases in classification accuracy over DDCCA-ND were small, indicating that the different decorrelation schemes have a minimal effect; the task-driven component is the main reason for the success of these methods. In particular, the classification accuracy with $n = 1000$ did better than the unsupervised method DCCA on $n = 10,000$. Further, DDCCA with $n = 300$ did better than linear methods on $n = 50,000$, showing the benefits of both the task-driven and deep components of the model.

The only DDCCA method that did not show as good results as the others is DDCCA-ED, which computes the linear CCA projection within the network. Computing the CCA projection directly rather than optimizing for the same ℓ_2 distance criterion with some method of decorrelation degraded performance.

Input Feature Dimension. In order to study robustness to HDLSS data, I manipulated the input feature dimension by adding extra features to both the left and right modalities. I increased the input feature dimension d by 2 or 3 times by adding additional features randomly sampled uniformly between 0 and 1. Figure 5.4b plots the cross-modal classification accuracy for each input dimension size for a training set size of $n = 1000$. RCCA and PLS-SVD showed an almost consistent accuracy across d. CCA, as expected, did not hold up to larger d. SoftCCA also proved to be very susceptible to large d. All other methods showed only a moderate decrease in performance for larger d.

Amongst the DDCCA methods, DDCCA-ED once again did not perform as well as the others for similar reasons as already discussed above. The other three produce very similar results, with DDCCA-SD giving a slight edge. All three methods are more robust than the others to HDLSS data as they make use of the task-driven component to find features that are both correlated and discriminative. Larger values of d will need to be tested to see if there is a point where DDCCA performance is comparable to RCCA and PLS-SVD.

Visualization. I also examined the CCA projections qualitatively by plotting them in 2D with t-distributed Stochastic Neighbor Embedding (t-SNE) [Van Der Maaten and Hinton, 2008]. Figure 5.5 shows the CCA projection of the left modality for each method. As expected, the task-driven methods all produce more clearly separated classes.

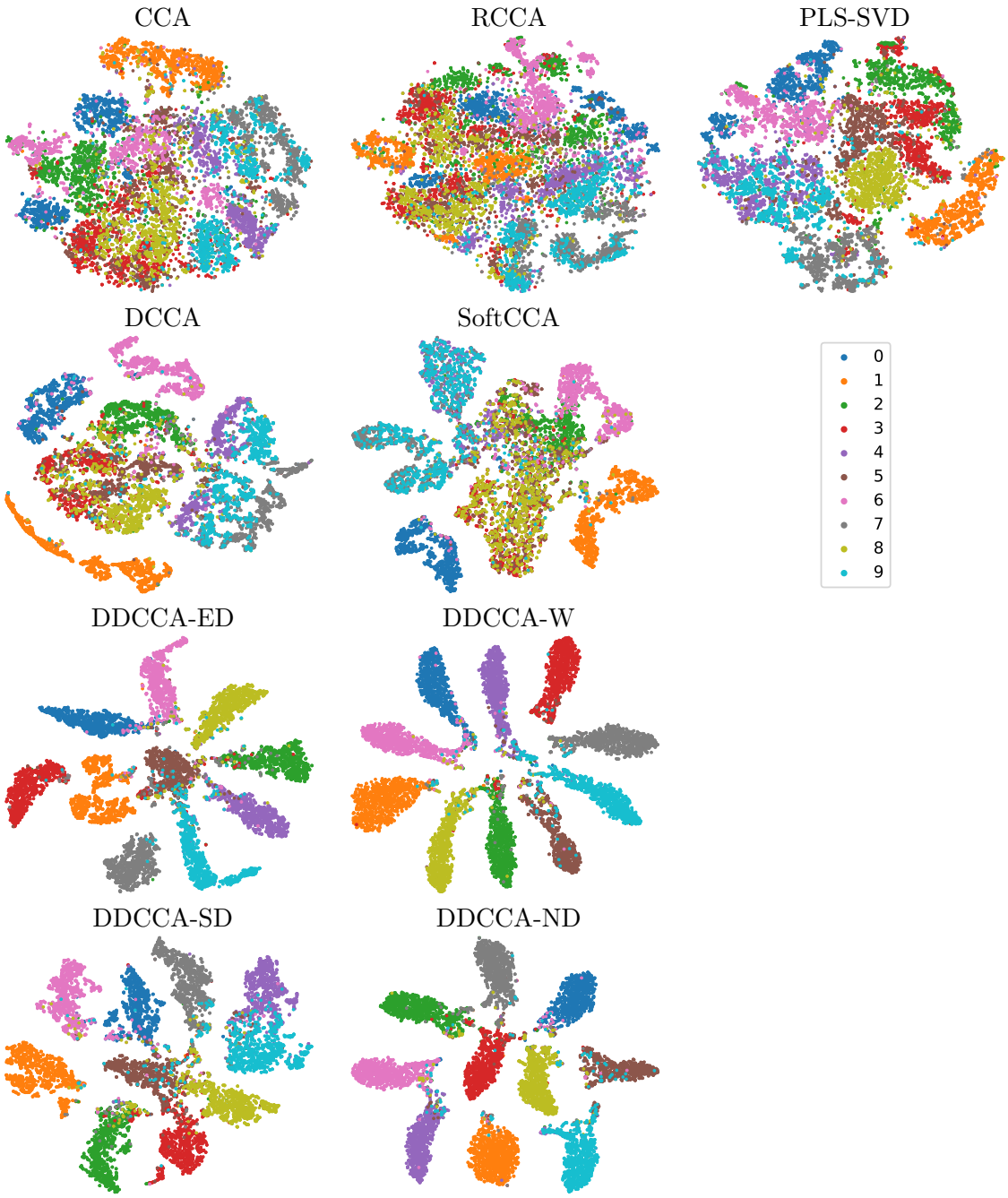

Figure 5.5: t-SNE plots for CCA methods on MNIST split. Each of the methods was used to compute projections for the left and right modalities (left and right sides of the images) using 10,000 training examples. The plots show a 2D visualization of the projection for the left modality computed with t-SNE with each digit colored differently. The samples for each digit are better clustered with the task-driven methods than with the unsupervised ones.

5.5.3 Cross-modal Classification on Real Data

Small Training Set: Carolina Breast Cancer Study (CBCS). This data set consists of image and genomic data for 1003 patient samples from the Carolina Breast Cancer Study, Phase 3 [Troester et al., 2018]. There are typically four TMA cores per patient. Image features were extracted with the CNN VGG16 [Simonyan and Zisserman, 2015] by taking the mean of the 512-dimensional output of the fourth set of convolutional layers across the tissue region and further averaging across all cores for the same patient. For gene expression, I used the set of 50 genes in the PAM50 array [Parker et al., 2009] or a larger set of 163 genes (referred to as GE163). The data set was randomly split into half for training, one quarter for validation, and one quarter for testing. Classification tasks included predicting Basal vs. non-Basal genomic subtype, estrogen receptor (ER) status positive vs. negative, and grade 1 vs. 3. Cross-modal classification accuracy is reported as the mean over four random training/validation/test splits.

Table 5.1 compares cross-modal methods. The cross-modal projection was computed using each method, following by training a linear SVM on the projection from the first modality. Testing was done by projecting the second modality and predicting its class using the SVM trained on the first modality.

While the left and right modalities on MNIST had a roughly equal ability to predict the digit class, this CBCS data set is much less symmetric. The ground truth genomic subtype was computed directly from PAM50, so it is expected that this modality is much better able to predict genomic subtype. ER was also better predicted from PAM50, while grade was assessed by a pathologist from histology images and therefore better predicted from image features.

The top performing method on CBCS was consistently DDCCA-W or DDCCA-SD. DDCCA-W was typically the better of the two, indicating that decorrelation by whitening is especially beneficial on this data set. These methods outperformed all unsupervised linear and deep CCA methods.

I also explored CCA methods qualitatively. Figure 5.6 shows t-SNE plots for all methods on both the image features and PAM50. As expected, the PAM50 data more clearly separated classes than the image features, at least for Basal vs. non-Basal and ER status. On both modalities, the DDCCA methods all produced a better separation of classes than the

	Train PAM50, Test Image		
Method	Basal	ER	Grade
CCA	0.732 (0.010)	0.637 (0.008)	0.741 (0.005)
RCCA	0.815 (0.008)	0.811 (0.003)	0.877 (0.010)
PLS-SVD	0.650 (0.016)	0.656 (0.003)	0.797 (0.010)
DCCA	0.787 (0.010)	0.785 (0.011)	0.867 (0.012)
SoftCCA	0.780 (0.010)	0.769 (0.014)	0.848 (0.015)
DDCCA-ED	0.802 (0.015)	0.803 (0.029)	0.852 (0.011)
DDCCA-W	**0.820 (0.008)**	**0.828 (0.006)**	**0.917 (0.019)**
DDCCA-SD	0.796 (0.004)	0.811 (0.004)	0.874 (0.019)
DDCCA-ND	0.766 (0.013)	0.805 (0.007)	0.878 (0.011)
	Train Image, Test PAM50		
Method	Basal	ER	Grade
CCA	0.943 (0.005)	0.869 (0.006)	0.828 (0.011)
RCCA	0.978 (0.003)	0.905 (0.008)	0.836 (0.009)
PLS-SVD	0.912 (0.018)	0.865 (0.005)	0.794 (0.012)
DCCA	0.976 (0.005)	0.874 (0.010)	0.854 (0.015)
SoftCCA	0.978 (0.004)	0.897 (0.010)	0.843 (0.007)
DDCCA-ED	0.971 (0.009)	0.889 (0.010)	0.833 (0.010)
DDCCA-W	**0.983 (0.006)**	**0.908 (0.009)**	0.845 (0.005)
DDCCA-SD	0.978 (0.005)	**0.908 (0.009)**	**0.848 (0.004)**
DDCCA-ND	0.975 (0.004)	0.896 (0.005)	0.817 (0.005)

Table 5.1: Cross-modal classification accuracy for CBCS. The standard error is in brackets. The CCA projection for each modality was first computed. A classifier was then trained on the projection from one modality, and the projection from the other modality was used at test time.

unsupervised CCA methods. The plot for SoftCCA is rather peculiar and will require further investigation. However, it does roughly order points from low-grade, non-Basal to high grade, non-Basal to Basal - an ordering that does make some sense. This "stringy" property was also seen on MNIST for some hyperpameter settings (not shown).

HDLSS Data: TCGA-BRCA. These experiments use genomic, epigenomic, and proteomic data from the breast cancer portion of TCGA. I used the following four modalities of data that are available for 466 tumors: gene expression (GE), miRNA expression, methylation, and Reverse Phase Protein Array (RPPA). While histology images are available for TCGA-BRCA, they are whole slide images rather than TMAs, bringing complications that are not yet solved by standard software and were not studied in this dissertation. Therefore, no imaging modalities were used for this data set. Each data type was obtained from a previous publication that analyzed these and other modalities [Network et al., 2012]. GE is the highest dimensional with 12,749 features, methylation has 574, miRNA has 802, and RPPA has 171.

The data set was randomly split into half for training, one quarter for validation, and one quarter for testing. Classification tasks included predicting Basal vs. non-Basal genomic subtype, estrogen receptor (ER) status positive vs. negative, and progesterone receptor (PR) status positive vs. negative. Cross-modal classification accuracy is reported as the mean over four random training/validation/test splits.

Table 5.2 displays the cross-modal classification accuracy for GE paired with each of the other modalities. DDCCA-W or DDCCA-SD was usually the top performer. When DDCCA was the top performer, it improved over RCCA by as much as 2.6%. The linear method RCCA was remarkably powerful on both CBCS and TCGA-BRCA. We can gain some insight by examining the DCCA and SoftCCA results that use an unsupervised method for deep CCA. DCCA and SoftCCA were typically inferior to RCCA, indicating that a deep version of CCA does not improve upon a linear method on this data set. Therefore, the increase in performance from DCCA and SoftCCA to DDCCA-W and DDCCA-SD comes from the task-driven component. The task-driven component acts as a regularizer and reduces overfitting.

While the proposed DDCCA methods were not always the top performer on TCGA-BRCA, the results do show that a deep method can be successful on HDLSS data. This indicates that

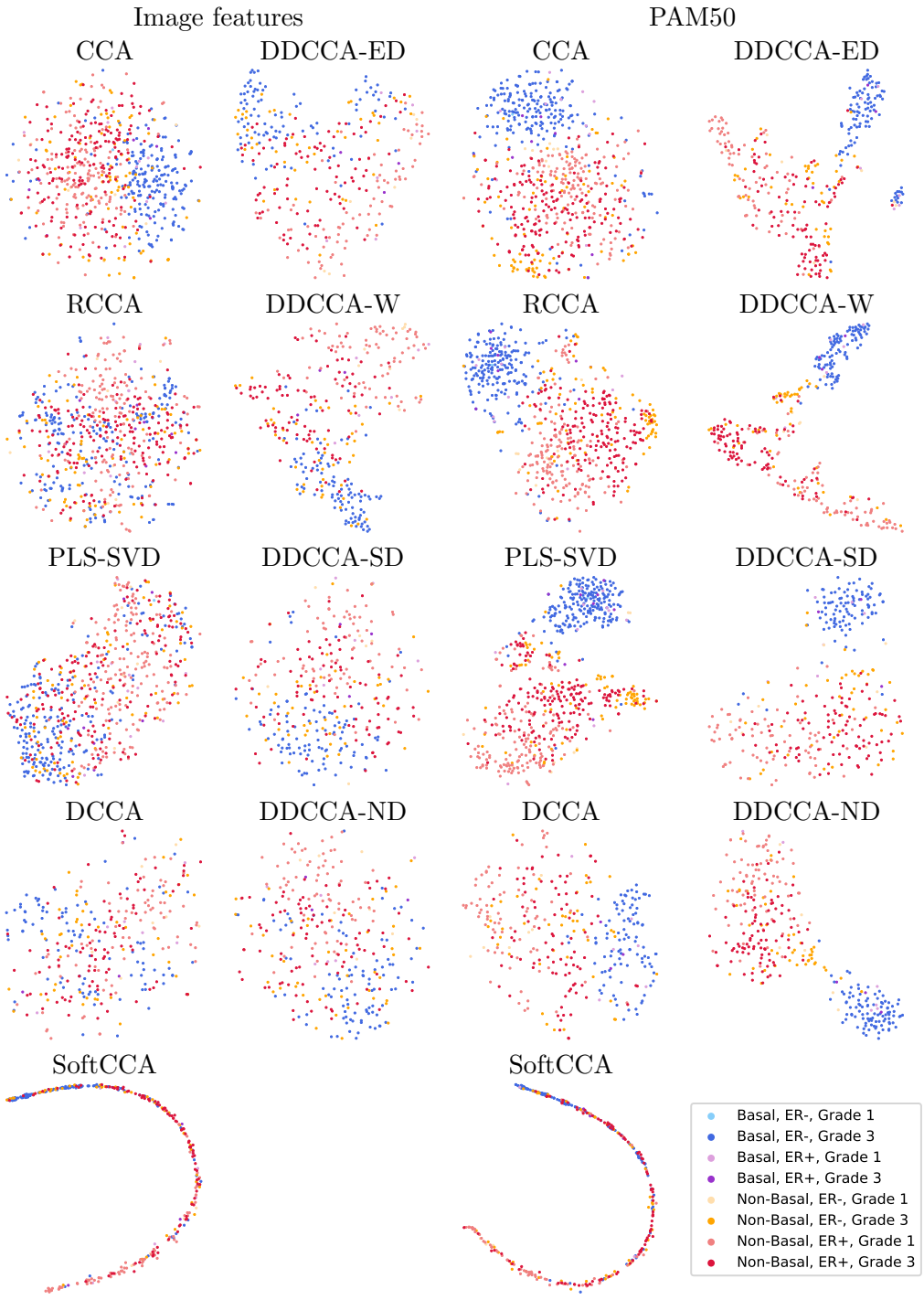

Figure 5.6: t-SNE plots for CCA methods on CBCS. Each of the methods was used to compute projections for image features and PAM50. The plots show a 2D visualization of the projection for each modality computed with t-SNE. Each point is colored according to its class. The classes are better separated with the task-driven methods than the unsupervised ones.

	Train GE, Test Methylation			Train Methylation, Test GE		
Method	Basal	ER	PR	Basal	ER	PR
CCA	0.913 (0.014)	0.837 (0.014)	0.731 (0.020)	0.880 (0.020)	0.835 (0.036)	0.784 (0.026)
RCCA	**0.978 (0.011)**	**0.919 (0.011)**	**0.834 (0.009)**	0.972 (0.006)	0.927 (0.010)	0.826 (0.006)
PLS-SVD	0.796 (0.016)	0.870 (0.016)	0.757 (0.018)	0.863 (0.011)	0.842 (0.018)	0.775 (0.015)
DCCA	0.963 (0.007)	0.885 (0.014)	0.775 (0.014)	0.983 (0.008)	0.899 (0.014)	0.815 (0.008)
SoftCCA	0.976 (0.004)	0.914 (0.006)	0.814 (0.014)	0.985 (0.004)	0.923 (0.008)	0.834 (0.012)
DDCCA-W	0.971 (0.006)	0.903 (0.016)	0.821 (0.012)	0.987 (0.005)	**0.930 (0.010)**	**0.837 (0.004)**
DDCCA-SD	0.971 (0.006)	0.916 (0.014)	0.826 (0.009)	**0.993 (0.004)**	0.921 (0.015)	**0.837 (0.007)**
	Train GE, Test miRNA			**Train miRNA, Test GE**		
Method	Basal	ER	PR	Basal	ER	PR
CCA	0.852 (0.010)	0.775 (0.031)	0.662 (0.024)	0.978 (0.007)	0.927 (0.008)	0.746 (0.021)
RCCA	0.946 (0.017)	0.901 (0.006)	0.790 (0.013)	0.976 (0.005)	0.923 (0.008)	0.826 (0.010)
PLS-SVD	0.841 (0.006)	0.819 (0.013)	0.768 (0.022)	0.885 (0.010)	0.859 (0.024)	0.788 (0.025)
DCCA	0.922 (0.013)	0.872 (0.012)	0.764 (0.017)	0.980 (0.009)	0.912 (0.0112)	0.788 (0.024)
SoftCCA	**0.959 (0.009)**	0.890 (0.013)	0.797 (0.011)	0.985 (0.002)	0.903 (0.006)	0.819 (0.013)
DDCCA-W	0.943 (0.009)	**0.914 (0.015)**	0.810 (0.020)	**0.991 (0.003)**	**0.932 (0.005)**	**0.843 (0.008)**
DDCCA-SD	**0.959 (0.010)**	0.912 (0.011)	**0.813 (0.019)**	0.985 (0.005)	0.921 (0.009)	0.823 (0.014)
	Train GE, Test RPPA			**Train RPPA, Test GE**		
Method	Basal	ER	PR	Basal	ER	PR
CCA	0.797 (0.029)	0.703 (0.021)	0.590 (0.005)	0.757 (0.013)	0.723 (0.021)	0.767 (0.025)
RCCA	**0.940 (0.011)**	**0.921 (0.017)**	**0.837 (0.018)**	0.966 (0.011)	0.910 (0.090)	0.817 (0.025)
PLS-SVD	0.857 (0.013)	0.814 (0.023)	0.776 (0.024)	0.891 (0.010)	0.852 (0.015)	0.794 (0.010)
DCCA	0.897 (0.021)	0.875 (0.022)	0.747 (0.025)	0.977 (0.004)	0.881 (0.009)	0.759 (0.025)
SoftCCA	0.934 (0.003)	0.872 (0.034)	0.805 (0.038)	0.983 (0.005)	0.899 (0.023)	0.805 (0.030)
DDCCA-W	0.931 (0.007)	0.904 (0.019)	0.805 (0.038)	**0.989 (0.006)**	0.904 (0.014)	0.831 (0.028)
DDCCA-SD	0.906 (0.011)	0.901 (0.018)	0.805 (0.026)	0.963 (0.012)	**0.919 (0.012)**	**0.843 (0.024)**

Table 5.2: Cross-modal classification accuracy for TCGA-BRCA. The standard error is in brackets. The CCA projection for each modality was first computed. A classifier was then trained on the projection from one modality and the projection from the other modality was used at test time.

replacing one of the modalities with a CNN operating on images could be successful even if the other modality is HDLSS data. Further, image data would provide an easy opportunity for data augmentation, likely improving the method further.

5.5.4 CCA for Regularization on CBCS

I further tested DDCCA as a method of regularization. If two modalities are available for training but only one at test time, the additional modality may help to regularize the model. I tested different methods of predicting genomic subtype, ER status, and grade when only images were available at test time. These methods include a) a linear SVM trained on image features, b) a DNN trained on image features, c) DDCCA-W trained on image features and PAM50, d) DDCCA-SD trained on image features and PAM50, e) DDCCA-W trained on image features and GE163, and f) DDCCA-SD trained on image features and GE163. Table 5.3 provides the

Method	Training data	Basal	ER	Grade
Linear SVM	Image only	0.785 (0.004)	0.838 (0.003)	0.897 (0.006)
DNN	Image only	0.796 (0.007)	**0.852 (0.008)**	0.907 (0.009)
DDCCA-W	Image+PAM50	0.827 (0.006)	0.839 (0.007)	**0.911 (0.012)**
DDCCA-SD	Image+PAM50	0.820 (0.010)	0.826 (0.009)	0.859 (0.021)
DDCCA-W	Image+GE163	**0.840 (0.010)**	0.838 (0.009)	0.910 (0.017)
DDCCA-SD	Image+GE163	0.812 (0.011)	0.815 (0.020)	0.891 (0.020)

Table 5.3: Classification accuracy for different methods to predict from images only at test time. Linear SVM and DNN were trained on only images. DDCCA-W and DDCCA-SD were used to regularize with PAM50 or GE163 during training. The standard error is in brackets.

classification accuracy for each method. While ER and grade only showed a small improvement beyond a linear SVM or DNN, or sometimes no improvement at all, genomic subtype Basal showed a much larger improvement of up to 5%. In particular, PAM50 and GE163 helped to regularize in both the DDCCA-W and DDCCA-SD models. This demonstrates that having additional information at training time can boost model performance at test time - dependent upon the additional data available and what is being predicted. Further, DDCCA-W with GE163 produced the best classification results for Basal vs. non-Basal from images at test time.

5.6 Discussion

The methods developed in this chapter add a task-driven component to deep CCA in such a way that it can be trained end-to-end. By making use of labeled data, they increase robustness to small training set sizes and HDLSS data.

All four methods were designed to allow small mini-batch training, the benefits of which were verified with an experiment. The methods differ in how they compute CCA within the network. Computing the CCA projection directly (DDCCA-ED) was shown to be overly restrictive. Applying orthogonality instead with whitening (DDCCA-W) or loosening it with soft decorrelation (DDCCA-SD) produced the best results. Removing the decorrelation piece entirely (DDCCA-ND) was only slightly inferior.

On the MNIST split data set, DDCCA showed large improvements in cross-modal classification accuracy; however, much smaller increases were observed for the CBCS and TCGA-BRCA data sets. In these cases, the linear method RCCA performed almost as well and sometimes

better than DDCCA. Although experiments on CBCS used a static set of image features and image features were not accessible for TCGA-BRCA, the DDCCA results indicate that a DNN solution is feasible, even with a small training set or HDLSS data. Thus, one of the modalities could be replaced with a CNN operating on images. This would also provide an opportunity for data augmentation. A larger variety of data sets still need to be explored to fully understand the benefits of DDCCA.

CCA can also be used as a regularization method when a second modality is available for training but not at test time. This strategy showed success on CBCS in predicting Basal vs. non-Basal genomic subtype. This regularization technique could be beneficial in fine-tuning a CNN on image data, perhaps by integrating with the methods of Chapter 4.

Finally, although DDCCA has been presented as a supervised method, it could also be implemented in a semi-supervised manner or even run when paired modalities are only available on a subset of the training set. This is possible because of the two components of the loss function: task-driven and ℓ_2 distance. When labels are not available for a particular training item, the task-driven loss can be ignored. When one modality or the other is missing, the ℓ_2 distance can be ignored.

CHAPTER 6: CONCLUSIONS

6.1 Summary of Contributions

This dissertation presented image and data analysis methods tackling three main areas: 1) representing histology images, 2) classifying heterogeneous images with multiple instance learning, and 3) integrating multimodal data. The contributions from this work are restated here with a discussion of how each was accomplished and the broader implications of each.

1) *Discriminative representations for histology images using dictionary learning or deep transfer learning. The dictionary learning method is task-driven to discover subtle differences between classes and hierarchical to capture architectural properties. The deep transfer learning method validates the use of pre-trained CNN features for discriminative tasks on non-RGB images.*

   While traditional approaches to representing images involved hand-crafted or hand-engineered features, my work took the approach of learning representations from the data. Dictionary learning formed this representation directly from image patches. The task-driven extension further informed the representation learning process by adding supervision through image labels.

   Dictionary learning can be used as a discriminative representation for classification of textures, objects, scenes, or many other types of images. It can also form a compact descriptor for image retrieval. While my patch-based application of task-driven dictionary learning was only moderately successful, this method would be better utilized when stronger labels are available - such as image-level labels for small images or fine-grained annotations for image patches. My patch-based application may still be beneficial when there is less heterogeneity present, for example in classifying textures. While I selected classification as the task-driven objective, other choices such as for regression or cluster-

ing can extend the applicability of this method to many other imaging and non-imaging domains.

I began to apply deep learning to histology by extracting features with a pre-trained CNN trained on ImageNet. While ImageNet and H&E histology are vastly different forms of data, the pre-trained CNN was diverse enough to provide meaningful features in discriminating histology images. Further, I used stain normalization and extracted the individual stain channels for use as input to the CNN. Even this non-RGB representation was shown to provide powerful features for classification.

Deep transfer learning provides an easy and powerful feature set without the computationally-intensive training of a full CNN. The power of this feature set comes from the space that the features span. I showed that features from a pre-trained CNN enable classification of H&E images; this technique is equally applicable to other tissue stains and, more generally, to other imaging modalities. These could include X-Ray, Computed Tomography, Magnetic Resonance, Ultrasound, fluorescence, and hyperspectral imaging. Three channels are likely not even a requirement; fewer channels could be applied by replicating channels, and more channels by applying the pre-trained CNN to different subsets of three channels.

2) *Multiple instance (MI) learning methods for handling large, heterogeneous images with an SVM on any type of feature set or with a CNN for end-to-end training. An iterative SVM-based method learns the latent instance labels given a particular assumption on instance label aggregation. Alternatively, a more general MI method uses the quantile function for pooling and learns how to aggregate instance predictions. This quantile method works with either an SVM or end-to-end training with a CNN. An MI augmentation technique is used while training the CNN and enables the exploration of single instance and MI learning on a continuous spectrum. Insight into both SVM- and CNN-based methods is provided by visualizing the predictions of each instance.*

The SVM-based methods performed binary classification into a positive or negative class. While the standard MI assumption used in most prior work states that a bag is positive if and only if at least one of its instances is positive, my work generalized this

assumption. My iterative SVM-based MI method began with an MI assumption that a given percentage of instances must be positive to assign a bag label of positive. The SVM classifier and latent instance labels were learned jointly to optimize bag classification performance, while enforcing the MI assumption at each iteration. Alternatively, my quantile aggregation method learned how much heterogeneity to expect in each class. This method assigned the bag label to each instance while training the SVM, and it aggregated instance predictions with the quantile function. The quantile aggregation method performed almost as well as the iterative MI method but without the need for the lengthy iterative process.

The CNN-based method brought MI learning to the multi-class case and enabled end-to-end training of the image representation and classifier. A CNN was applied fully convolutionally with an MI aggregation layer on top to form the bag prediction. Quantile aggregation was used once again to form a more complete description of the distribution of instance predictions and learned how much heterogeneity to expect for each class. Each of these additions to the CNN model maintained the ability to backpropagate errors for end-to-end training. MI augmentation during training provided a mechanism to study single instance and multiple instance learning on a continuous spectrum, demonstrating that MI learning is essential on heterogeneous images.

MI learning provides a solution when data annotations are weak - labels are applied to a group of items instead of each individual item. By generalizing the instance aggregation method and, further, by learning how to aggregate instances, MI learning can now be applied to a much wider range of applications - in particular, those for which an appropriate aggregation function is not known. These methods are applicable whenever labels are weakly applied to sets of instances, including large heterogeneous images. Other heterogeneous diseases that these methods could be applied to include chronic obstructive pulmonary disease and Alzheimer's disease; any imaging modality may be used. Beyond medical applications, these techniques can be applied whenever there are weak annotations for a concept due to the time or expense required for creating finer-grained annotations or due to the inability of humans to perform such an annotation task.

3) *A set of multimodal methods to find a shared space that is also discriminative. This set of deep CCA models can be used for cross-modal classification and in gaining insight into the shared components of two modalities. They bring the CCA projection into the network itself in different ways, enabling end-to-end training to optimize both the correlation between modalities and the task-driven goal.*

While existing methods for deep CCA optimize the sum correlation between a pair of modalities, they do not compute the projection to the shared space within the network but as a post-processing step. In order to use the projection in end-to-end training with a task-driven goal, I developed four different methods for incorporating a CCA-style projection within the DNN, each handling the orthogonality constraints of the original CCA model in a different way. These task-driven deep CCA methods can use any task-driven goal such as for classification or clustering, are shown to outperform non-task-driven alternatives, and are much more robust to HDLSS data.

The robustness properties of the methods that I proposed make them especially suitable for medical applications in which patient samples are limited and input dimensions can be high in many modalities such as genomics and proteomics. Many other data types could be explored, including clinical data and other imaging modalities such as X-Ray, Computed Tomography, or Magnetic Resonance. Multimodal data is also prevalent across a variety of applications: audio and video, images and text, parallel texts in different languages, or even different feature representations for the same data. Text representations using some measure of word frequency are also very high dimensional, which may be particularly well-suited to these methods. My proposed task-driven deep CCA methods can also be applied in a semi-supervised manner when labels are only available for some samples and, further, when a parallel modality is missing for some samples but a label is provided.

Beyond classification tasks, CCA-based methods also provide insight into shared properties of data. For cancer, they could enable the study of how much of the variation in genomic data is exhibited in imaging modalities such as histology. Clustering to find subtypes of a disease could be done in the shared projection space instead of with a single

modality. Any task-driven goal that can be used in a deep network can also be used in these task-driven multimodal methods.

4) *Techniques for deep learning on problems traditionally viewed as "small data." Solutions in this regime include deep transfer learning, multiple instance learning on large images, multi-task learning, and appropriate regularization.*

The applications addressed in this dissertation have on the order of 1000 labeled examples, not the millions typically used in training a DNN. Using a pre-trained CNN for feature extraction provided the simplest adaption to smaller data sets, even on disparate image types. In order to fine-tune a large CNN with weak labels on these large images, MI learning was necessary to handle image heterogeneity. With this technique even today's largest CNNs can be fine-tuned on an image data set with fewer than 1000 patients. The task-driven deep CCA models go a step further to HDLSS data and are successful because they downsize the input before applying a CCA-style operation. The techniques of whitening and soft decorrelation are used to decorrelate features within each modality and are actually forms of regularization. In both MI and multimodal learning, a multi-task setup to predict five different image properties provided additional regularization.

Deep learning has had limited study on smaller data sets because it best lends itself to the use of large training sets to learn powerful models with many parameters. However, non-linear representations also benefit smaller data sets and were shown to be successful with MI and multimodal learning. Small training set size is a common trait for many medical applications due to the time and expensive of gathering both patient data and expert annotations. When learning models for challenging problems on small data sets, any additional information available can strengthen the model, including additional modalities and additional class labels. This multi-task framework exemplifies the importance of regularization by learning a shared representation across tasks. This technique is broadly applicable but is especially important on small data sets.

Further contributions were made to the application area of breast cancer research. Each is discussed here.

A) *Methods to capture biologically-relevant features by operating on the H&E stain intensities extracted from histology images. These methods do not rely on hand-crafted features and are shown to produce more accurate predictions. The feature learning methods are also easily transferable to other cancer types.*

While prior work has applied stain normalization to H&E histology images, I took the extra step of using the extracted stain channels of hematoxylin, eosin, and residual, rather than simply the normalized RGB image, as input when forming image representations. This pre-processing step aided representation learning by focusing on the nuclei content that was stained blue by hematoxylin and the cytoplasm content stained pink or red by eosin. Further, the features learned by all methods in this dissertation are shown to be discriminative for biologically important classifications such as histologic subtype, grade, estrogen receptor status, and genomic subtype. Because these methods are trained on the given data with only patient-level labels, they are easily transferable to other tissue types or other imaging modalities.

The success of the methods that I propose demonstrates the power of representation learning on a task typically handled by a well-trained expert. Given sufficient data, computers can learn which properties of an image are most helpful for a particular classification task. Some of these characteristics may be complex and abstract. However, these methods come with a large downside: the lack of interpretability of individual features. This is an active area of research for histology and, more broadly, in deep learning as a whole.

B) *A low cost and repeatable method for predicting histopathological, molecular, and genomic properties of tumors from H&E histology. Experimental validation shows that the classification accuracies achieved are comparable to the inter-rater agreement of pathologists and of alternative ways of assessing tumor properties. Further, these methods showed success on predicting molecular and genomic properties from H&E histology - something not previously known to be possible from H&E alone.*

Molecular methods such as immunohistochemical staining to determine ER status and PAM50 subtyping are expensive and not routinely performed. Grading is routinely done

by a pathologist but can be highly variable for intermediate grade tumors. Computational methods are more repeatable. The image classification methods in this dissertation have shown a high classification accuracy for histologic subtype, grade, ER status, genomic subtype, and risk of recurrence score. Statistical validation done in joint work shows that these methods may provide a viable alternative to or screening method for expensive molecular tests [Couture et al., 2018b].

While grade and histologic subtype are routinely assessed visually by a pathologist, ER status and genomic subtype were not previously known to be predictable from H&E histology alone. Pathologists are very experienced in identifying tumors on large whole slide images and assessing grade but are still limited by their human ability to process complex information. It was only through automated computational methods that I was able to show that these molecular and genomic classes are distinguishable from H&E alone.

PAM50 subtyping using gene expression has already shown utility in guiding treatment decisions [Parker et al., 2009]. With further refinement the methods that I propose on histology may provide a viable alternative. However, some additional challenges will need to be studied: the lack of standardization across labs for tissue preparation and staining protocols, differences in image acquisition procedures, and the selection of tumor tissue from larger sections of tissue. While stain normalization was used as pre-processing in this work, it may not be needed. If sufficient variety is present in the training data and additional variability is generated with data augmentation, countering these effects may not be necessary, particularly if a large training set is provided. Finding tumor in whole slides images remains an active area of research.

Given sufficient labeled data, computers can learn concepts much more complex than even the best trained human experts. While molecular subtypes of breast cancer is one example, machine learning, and deep learning in particular, can provide suitable methods for many complex tasks from assessing prognosis of cancer [Katzman et al., 2016; Wang et al., 2017] to predicting earthquake aftershocks [DeVries et al., 2018]. The key is a sufficiently large set of labeled data - the more the better. Additional labeled data would

likely produce further improvements in the classification tasks tackled in this dissertation.

C) *A mechanism to find predicted tumor heterogeneity from H&E histology. While genomic subtyping methods assess a very small region of tumor, imaging provides a spatial view. Predicting tumor subtype from smaller regions across the image provides a view of heterogeneity. Future work will be necessary to validate the predicted tumor heterogeneity and assess its association with patient outcome.*

Genomic subtyping methods are generally only performed on one small sample from each tumor, eliminating any opportunity to assess intra-tumor heterogeneity. Spatial information remains, however, in histology images. While processing whole slide images would provide a more complete view of heterogeneity, the four cores extracted for use in the TMAs on the CBCS data set provide a first insight. By capturing heterogeneity while training the predictive models and, further, by making class predictions for each core and within cores, finding and quantifying tumor heterogeneity is now possible.

The implications of intra-tumor heterogeneity are not yet understood. At minimum, it imposes a challenge for diagnosis and subtyping [McGranahan and Swanton, 2015]. It likely also increases the complexity of treatment decisions [Hiley and Swanton, 2014] and may cause tumors to be more aggressive and resistant to treatment [Natrajan et al., 2016]. A great deal more study of intra-tumor heterogeneity is needed, for which my methods to find predicted heterogeneity can provide a unique view.

These methods could also provide biological insights into other heterogeneous diseases. More broadly, intra-class heterogeneity is present in many other classification problems from environmental to economic. Any time a label is applied to a group of cells, people, households, organisms, or other potentially diverse components, some amount of intra-class heterogeneity is present. The methods that I propose could provide insight into the amount of heterogeneity in a data set and, from that, a better understanding of the data itself.

Finally I revisit the thesis statement presented in Chapter 1:

Learned representations for histology images of tissue can capture both intra- and inter-tumor heterogeneity, enabling discriminative models for tumor properties. Combining these

image features with data from other modalities such as genomics in a task-driven model can provide insight into the shared tumor properties and further improve predictions. These computational techniques using discriminative features can provide a lower cost and more repeatable alternative to molecular methods and insight into tumor progression and heterogeneity.

Image representations were initially explored in Chapter 2 and further developed within an MI framework in Chapters 3 and 4. These methods learned discriminative properties of breast tumors and, through MI learning, captured tumor heterogeneity. The discriminative feature learning methods developed in this dissertation were shown to produce a higher classification accuracy than previous hand-crafted and hand-engineered techniques. Statistical validation was provided by predicting breast tumor grade, histologic subtype, estrogen receptor status, genomic subtype, and risk of recurrence score, demonstrating their utility as an alternative to or screening method for expensive molecular tests. The MI framework presented not only accounts for intra-tumor heterogeneity but also enables its quantification. Future work will be needed to validate the predicted heterogeneity and its association with patient outcome.

The set of multimodal methods developed in Chapter 5 learn a projection to a shared space that is also discriminative, producing a large improvement in cross-modal classification accuracy on some data sets. Further, some of these methods were shown to be more robust to small training set sizes and HDLSS data than unsupervised CCA-based methods. These deep CCA methods can also be used to regularize a model when both modalities are available for training, but only one at test time, and was shown to be beneficial with breast tumor images and genomic data.

6.2 Software and Data Availability

Data. The main data set used throughout this dissertation is the Carolina Breast Cancer Study (CBCS), Phase 3 [Parker et al., 2009]. Further details on this data set and instructions for gaining access are available here: http://cbcs.web.unc.edu/for-researchers/.

Code. Source code for the proposed methods is available upon request. Source code for Chapter 5 may also be available publicly on GitHub at a later date.

6.3 Future Work

While the methods presented in this dissertation take new strides in analyzing histology images of tumor tissue, there are numerous directions for future study.

6.3.1 Training a CNN for Histopathology

Additional Data Augmentation. In training a CNN for histology in Chapter 4, data augmentation by cropping was used and adapted to form the MI augmentation technique. Data augmentation by random mirroring and rotations was also included in the training process. Further data augmentation that has previously been shown beneficial includes small changes in scaling and color distortions [Lafarge et al., 2017].

Semi-supervised Learning with a Generative Adversarial Network. The most recent success of deep learning is in the area of supervised learning and is made possible by data sets with millions of labeled images [Krizhevsky et al., 2012; Szegedy et al., 2015; Simonyan and Zisserman, 2015]. While unsupervised feature learning is an active area of research [Le et al., 2012b; Caron et al., 2018], its use in a classification model does not yet compare with supervised feature learning. This is particularly unfortunate in medical applications for which gathering that much finely-annotated data is cost prohibitive. In the case of histology, there is a great deal of unlabeled data available, making unsupervised or semi-supervised feature learning attractive.

While auto-encoders have shown some success on histology [Chang et al., 2015; Arevalo et al., 2015], the more recent Generative Adversarial Network (GAN) has rapidly become the chosen method for unsupervised feature learning on images [Goodfellow et al., 2014; Radford et al., 2015]. GANs generate data without explicitly modeling the probability density function. After training on unlabeled data, they can then be used as a feature extractor for supervised tasks. They have already been used for synthesizing H&E histology images [Hou et al., 2017] and H&E stain normalization [Bentaieb and Hamarneh, 2018; Shaban et al., 2018]. GANs could be further adapted to take advantage of the wealth of unlabeled histology data combined with labeled data using semi-supervised learning [Springenberg, 2015; Odena, 2016].

6.3.2 Multimodal Deep Learning

Shared and Individual Representations. The methods developed in Chapter 5 learn features that are shared between a pair of modalities. While this is essential for cross-modal classification, further insight into the data can be gained from also learning features that are unique to each modality. Dictionary learning-type methods have been applied in extracting these shared and individual components, with a focus on reconstructing the input data from the encoded representation [Ray et al., 2014b; Lock et al., 2013; Zhou et al., 2015]. Adapting such techniques to deep learning could take the form of an auto-encoder with a reconstruction objective or by maintaining the task-driven focus while learning individual representations that are orthogonal to the shared ones.

Sparse CCA. The challenges of adapting CCA-based models to HDLSS data were discussed in Chapter 5. While the deep methods presented in that chapter were shown to be successful on HDLSS data, other techniques can further improve performance in this difficult setting. Sparse CCA adds a sparse regularizer to minimize the number of input features used [Hardoon and Shawe-Taylor, 2011]. While this method has currently only been applied to linear CCA, it could be applied to the deep models developed in Chapter 5.

Data Augmentation. The methods developed in Chapter 5 were tested without data augmentation. Data augmentation techniques for images were already discussed above but could also be important for multimodal models, particularly on small training sets. While adding random noise is a first simple strategy, other options may be specific to the type of data, just as scaling and color distortions are specific to image data.

Multimodal Data for Training a CNN. Chapter 4 showed success in fine-tuning a large CNN to predict tumor properties like subtype and grade. While the multimodal techniques in Chapter 5 have only been applied to features extracted with a pre-trained CNN, a CCA-type objective could provide additional regularization or an alternative objective in fine-tuning a CNN. Image data also provides an easy opportunity for data augmentation.

6.3.3 Deep Learning Data Challenges

Ordinal Prediction. The end-to-end training of a CNN in Chapter 4 included both binary and multi-class tasks. However, some of these tasks, such as grade and risk of recurrence score, should not be represented as discrete classes but as a continuous spectrum. While performing this task as regression instead of classification is one possible solution, alternative solutions using multiple binary SVMs exist [Li and Lin, 2007; Chu and Keerthi, 2007]. A multi-task ordinal regression method for deep learning may also prove effective.

Deep Learning with Imbalanced Classes. Learning a classifier with imbalanced data presents challenges due to the skewed distributions of binary tasks. Learning with imbalanced data is an open area of research [Krawczyk, 2016]. Chapter 3 included a sample weighting method to boost the classification accuracy of under-represented classes, namely by weighting inversely proportional to a secondary class such as grade. For CNNs, data augmentation could boost under-represented classes using additional sampling. Alternative approaches provide a more efficient solution for deep learning [Huang et al., 2016; Khan et al., 2017].

Model Interpretation and Visualization. Initial attempts at interpretation in this dissertation involved heatmaps identifying which regions of the image are most associated with each class. Further insight could be gained by related techniques like saliency mapping [Simonyan et al., 2013] and layer-wise relevance propagation [Bach et al., 2015]. While these methods all provide insight into which regions of the image are most relevant, other approaches could provide visualizations [Zeiler and Fergus, 2014] or meaningful labels [Zhou et al., 2017] for individual features. Further insights could be gained by correlating network features or linear combinations thereof with hand-crafted cell morphology features.

Efficient Processing of Large Images on the GPU. GPU technology has pushed deep learning progress forward in leaps and bounds. While GPU memory has grown significantly, the size of image that can be processed during end-to-end model training is still limited, particularly for larger CNNs. Clearly the MI method presented in Chapter 4 will not be feasible on whole slide images due to GPU memory constraints; alternative MI solutions have been developed for

this situation [Hou et al., 2016]. By decreasing the batch size, I was able to process a single core image at once while training the large CNN AlexNet; however, even a full core image cannot be processed with larger CNNs like VGG16. Strategies involving multiple GPUs [Abadi et al., 2016] or other synchronization methods may provide assistance in this area.

6.3.4 Cancer Research

Region of Interest Selection from Whole Slides. This dissertation has focused on TMA images in which a pathologist selected core regions that are largely tumor. While four cores are selected from each tumor, analysis of the whole slide could provide additional information on tumor properties, particularly with regard to heterogeneity. Tumor detection and region of interest selection is currently a research area of itself due to the challenges of large image size, diverse image appearance, and the local image region annotations needed [Wang et al., 2016a; Cruz-Roa et al., 2017; Kong et al., 2018; BenTaieb and Hamarneh, 2018].

Tissue Segmentation. Individual cores in a TMA are hand-selected and punched from the tumor. Although they contain mostly tumor tissue, some adjacent stroma or adipose tissue can be present. Segmenting the epithelial tumor tissue from the rest would enable the capture of properties separately from each tissue type. Both Yuan et al. [Yuan et al., 2012] and Beck et al. [Beck et al., 2011] have found that stromal cell morphology is more predictive of prognosis than the appearance of tumor cells, highlighting the importance of properties of surrounding tissue and not focusing exclusively on the tumor itself. Future work in predicting prognosis would likely benefit from such a segmentation.

Outcome Prediction. One of the original goals of this project was to predict recurrence or survival from histology images. While recurrence data is available for the CBCS, Phase 3 data set, the length of follow up is still limited. Larger public data sets like The Cancer Genome Atlas include a much longer follow up with both recurrence and death information available; however, this data set uses whole slide images, bringing with it the challenges already mentioned. Future work addressing the processing of whole slides or accessing data with a longer follow up for TMA images will be needed to study outcome prediction.

Outcome prediction can be approached as a binary task (event occurred prior to N years or it did not); however, this sets up a very imbalanced classification problem. Techniques in machine learning for imbalanced data may alleviate this concern [Krawczyk, 2016]. Outcome prediction models like the Cox Proportional Hazard [Cox, 1992] or multi-task survival analysis [Li et al., 2016] are more suitable for right censored data. Initial attempts with such a model using image features indicated that a multi-task approach might be needed so that lower level features are shared by predictors for other known labels such as genomic subtype or grade.

Validation of Heterogeneity. Chapters 3 and 4 presented means for capturing and quantifying tumor heterogeneity. While the methods presented can predict the presence of heterogeneity, future work is needed to validate its presence. Heterogeneity of receptor status such as ER could be validated with IHC stains of adjacent slices. Validation of genomic subtype heterogeneity, however, would require processing of multiple core samples, rather than the single one typically done on most data sets. In the absence of data for validating the presence of heterogeneity, quantitative measures may still be of importance as their association with other known variables can be tested, such as their ability to predict patient outcome. The study of intra-tumor heterogeneity may lead to future insights in cancer progression [Alizadeh et al., 2015; McGranahan and Swanton, 2015].

Biological Insights. While I have shown that molecular and genomic subtypes are distinguishable from H&E histology though computational methods, it is not clear what properties distinguish them. The differences are likely subtle, as they are not visually apparent to pathologists. As discussed above, interpreting the abstract features from a CNN is an open research topic. Future work in this area may enable the discovery of what visual properties distinguish ER status and genomic types in H&E images. Such insights may provide a teaching opportunity for pathologists or biological insights into tumor subtypes.

Further insights may be gained by the multimodal methods in Chapter 5. In particular, the deep CCA models could indicate which genes manifest visually in the H&E images and which do not. This could be done by adding a sparse regularizer to minimize the number of input features, as discussed above.

Application to Other Cancer and Disease Types. While this dissertation has focused on breast cancer, the methods developed are generic. They can and should be tested on other tissue types, other cancer types, other diseases, and other types of heterogeneous and multi-modal data.

6.4 Closing Remarks

Some of the methods developed early in this dissertation have already shown utility for breast cancer as an alternative to or screening process for expensive molecular tests. Improvements in histology classification in subsequent chapters brought greater prediction accuracy and insights into tumor heterogeneity. Integration of genomic data has begun to show new insights into tumor classification schemes. While the algorithms developed herein were tested on breast tumor histology, my hope is that they will be further refined and applied to many other cancer and disease types - from improving treatment decisions to better understanding the disease itself. Multiple instance and multimodal methods are also extendable to aid discriminability on many forms of heterogeneous data.

APPENDIX A: EXPERIMENTAL VALIDATION OF BREAST TUMOR HISTOLOGY CLASSIFICATION[1]

A.1 Introduction

Image-based features of breast cancers have an important role in clinical prognostics. For example, tumor grade is strongly associated with survivorship, even among tumors with other favorable prognostic features such as estrogen receptor positivity [Dunnwald et al., 2007]. However, major advances in prognostication over the past decade have relied predominantly on molecular methods [Parker et al., 2009; Sparano and Paik, 2008; Carlson and Roth, 2013]. These methods are costly and are not routinely performed on all clinical patients who could benefit from advanced molecular tests. Methods are needed for identifying patients who are likely to benefit from further molecular testing.

We hypothesized that a deep learning method for image analysis could be applied to classify H&E-stained breast tumor tissue microarray (TMA) images with respect to histologic and molecular features. We used TMA images from the population-based Carolina Breast Cancer Study Phase 3 (2008-2013) to perform deep learning-based image analysis aimed at capturing larger scale and more complex properties including tumor grade, histologic subtype, Estrogen Receptor (ER) status, intrinsic breast cancer subtype and a risk of recurrence score (ROR-PT) [Parker et al., 2009].

A.2 Methods

The classification model used in this work is based off the SIL-quantile method detailed in Section 3.5. Images were stain normalized using the method by Niethammer et al. [Niethammer et al., 2010]. Image features were extracted using the output from the fourth set of convolutional layers from pre-trained VGG16 [Simonyan and Zisserman, 2015].

Image regions were generated as 800×800 pixel regions in the training images, with the mean of each CNN feature computed over the region. A linear SVM calibrated with isotonic

[1]The work presented in this appendix was published in npj Breast Cancer and was joint work with Lindsay Williams [Couture et al., 2018b]. I developed the classification methods and ran them on the data set to produce class predictions for each core and each sample. Lindsay Williams performed the statistical analysis from these class predictions. This appendix is a reproduction of the original article but with Sections A.1 and A.2 modified and condensed as much of the methods were already described in Chapter 3.

regression [Zadrozny and Elkan, 2002] was used to predict the probability for each region. Isotonic regression fits a piecewise-constant non-decreasing function, transforming the distance from the separating hyperplane learned by the SVM to a probability that an image region belongs to each class. This assumes that the SVM can rank image regions accurately and only needs the distances converted to probabilities. Each image region was labeled with the class of the tumor from which it belongs. The data for model fitting and calibration must be disjoint, so cross-validation was used to split the training instances into five equal-sized groups, where four were used for training and the remaining for calibration/validation (the test set remains untouched). For each fold, an SVM was learned on the training set, and calibration was learned on the calibration set with isotonic regression, thus forming an ensemble. An ensemble of size five was selected to balance the desirability of a large training set, a reasonably sized validation set, and the simultaneous desirability of limiting the computation time. Predictions on the test set were made by averaging probabilities from the five models. This ensemble method also helped to soften any noise in the predictions caused by incorrect image region labels due to heterogeneity.

Predictions for tumors were made by first forming a quantile function of the calibrated SVM ensemble predictions for the image regions using 16 equally spaced quantiles from images in the training set. The quantiles of the training images were used to train another linear SVM to predict the class label for the whole tumor, with sigmoid calibration transforming the SVM output into probabilities. This method allowed predictions to be made for individual image regions, while also aggregating to overall tumor predictions.

Sample weighting by grade (Section 3.6) was used to reduce the leverage of grade in predicting ER status and intrinsic subtype. Sample weighting was applied using weights inversely proportional to the number of samples in the group; that is, low grade class 1, low grade class 2, high grade class 1, and high grade class 2 were each weighted equally.

At test time, 800×800 pixel overlapping regions with a stride of 400 pixels were used as image regions from each TMA spot that is typically 2500 pixels in diameter. Only image regions containing at least 50% tissue within the core image field of view (i.e. 50% tissue, 50% glass) were used. The calibrated SVM ensemble predicted the class of each image region by assigning a probability of belonging to one of two classes (tumor grade 1 or 3, ER+ or ER-, Basal-like

or non-Basal-like subtype, ductal or lobular histologic subtype, and low-medium or high ROR-PT). The probabilities computed on the image regions from all cores were aggregated into a quantile function, and the second SVM was used to predict the class for the whole tumor.

A.3 Results

Sample Set. The training and test sets were both comprised of participants from the Carolina Breast Cancer Study (CBCS), Phase 3 (2008-2013) [Troester et al., 2018]. The training and test sets were formed by a random partition of the data. The total number of patients available for the training and test set from CBCS3 was 1,203. These patients were divided into a group of 2/3 (n=802) for the training set and 1/3 (401) for the test set. Of the 802 patients available for the training set, 571 had H&E images and biomarker data available for contribution to the training set. Of the 401 patients eligible for the test set, 288 had H&E images and biomarker data available. Patients in the final training and test sets had information for tumor grade and histologic subtype, determined via centralized breast pathologist review within CBCS, along with biomarker data for ER status, PAM50 intrinsic breast cancer subtype, and risk of recurrence (ROR-PT) where noted. The H&E images were taken from tissue microarrays constructed with 1-4 1mm cores for each patient, resulting in 932 core images for the test set analysis presented here. ER status for each TMA core was determined using a digital algorithm as described by Allott et al. (2015) [Allott et al., 2015] and was defined using a $\geq$10% positivity cut point for immunohistochemistry staining. There were no significant differences between the training and the test sets concerning patient or tumor characteristics (Table A.1).

Grade. Across multiple 1.0-mm cores per patient, the probability of a tumor being classified as high grade by image analysis was calculated; Figure A.1 shows that a bimodal distribution of probabilities was observed. By establishing a cut point at >0.80, high grade tumors were detected with an accuracy of 82% in the test set (kappa 0.64) (Figures A.1 and A.2, Table A.2). Considering low/intermediate as a group, the percent agreement with pathologist-classified tumor grade was slightly lower than the percent agreement between two breast pathologists who independently reviewed these same patients (overall 89%, kappa 0.78). Tumors with pathologist-defined intermediate grade were more likely to be misclassified as high grade tu-

	Training set (N=571) N (%1)	Test set (N=288) N (%1)	Chi-square p-value
Age			
≤50 years	280 (29.6)	133 (28.0)	0.64
>50 years	291 (70.4)	155 (72.0)	
Race			
White	298 (79.0)	150 (78.7)	0.90
African-American	272 (21.0)	138 (21.3)	
missing	1		
Grade			
Low-Intermediate	330 (65.8)	162 (66.5)	0.85
High	240 (34.2)	125 (33.5)	
missing	1	1	
Stage			
I, II	485 (86.4)	259 (90.2)	0.17
III, IV	85 (13.6)	29 (9.8)	
missing	1		
Node Status			
Negative	354 (65.2)	191 (69.1)	0.35
Positive	214 (34.8)	97 (30.9)	
missing	3		
Tumor Size			
≤2cm	334 (62.5)	174 (67.2)	0.26
>2cm	235 (37.5)	114 (32.8)	
missing	2		
ER Status			
Negative	164 (24.9)	91 (23.1)	0.62
Positive	405 (75.1)	197 (76.9)	
missing	2		
PAM50 Subtype			
Luminal A	149 (46.1)	74 (47.1)	0.27
Luminal B	78 (18.2)	33 (20.9)	
Basal-like	92 (20.9)	49 (21.6)	
HER2	46 (11.9)	15 (5.9)	
Normal-like	9 (2.9)	9 (4.5)	
missing	197	108	

Table A.1: Patient and tumor characteristics for the Image Analysis training and test set.

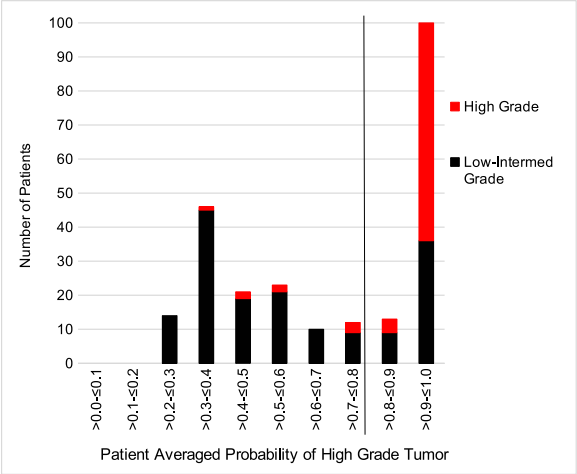

Figure A.1: Histogram for probability of high grade tumor by image analysis according to proportion of pathologist-classified low-intermediate (black) or high grade (red) in the test set. A cut point of >0.80 was selected.

mors by image analysis (37%) while only 7% of low-grade tumors were misclassified (results not shown). When comparing the misclassification of intermediate- and low-grade tumors as high grade between two pathologists in a subset of CBCS tumors, errors in classification of intermediate grade tumors as high grade tumors occurred <10% of the time and never occurred for low grade tumors (results not shown).

Molecular Characteristics. Image analysis accuracy for predicting molecular characteristics was also high. Accuracy for ER status was 84% (kappa 0.64) and both sensitivity (88%) and specificity (76%) were high (Table A.3). However, tumor grade is strongly associated with ER status in most patient populations, and we were interested in increasing accuracy among patients with low-to-intermediate grade tumors where genomic testing is most likely to influence patient care. Thus, we also employed a training strategy that weighted samples to ensure that low and intermediate grade distributions were similar between ER positive and ER negative tumors. This reduced accuracy among high grade tumors (from 77% to 75%) and decreased accuracy among low-intermediate grade tumors (from 91 to 84% accuracy). Using the same weighting strategy, we trained a classifier to predict Basal-like vs. non-Basal-like (Luminal A, Luminal B, HER2, Normal-like combined) PAM50 subtype (Table A.4). The classifier had an overall accuracy of 77% but an accuracy of 85% among low-intermediate grade tumors and 70%

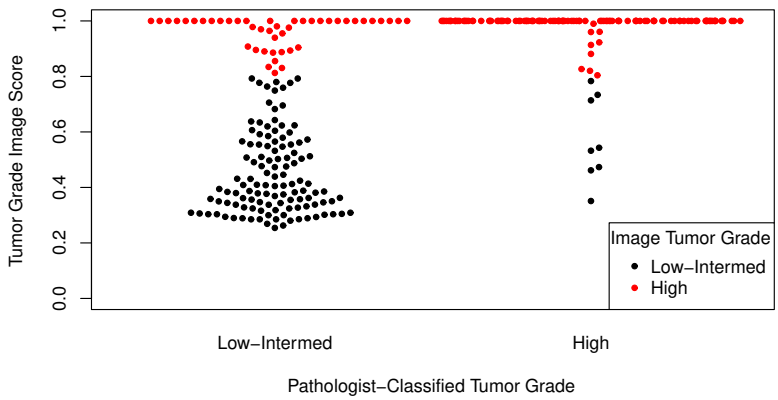

Figure A.2: Bee Swarm plot displaying pathologist classification of tumor grade as a function of the image grade score in the test set. Points within each grade group are adjusted horizontally to avoid overlap. The black dots indicate image analysis classified low-intermediate tumor grade, and the red dots indicate image analysis classified high grade tumors.

Pathologist Agreement on Tumor Grade Classification[1] (n=242)		
	Pathologist 2	
Pathologist 1	Low-Intermediate Grade	High Grade
Low-Intermediate Grade	113	23
High Grade	4	102
	% Agreement	89
	kappa	0.78
	(95% CI)	(0.70-0.86)
Image Analysis agreement with Pathologist Tumor Grade Classification[2] (n=288)		
	Clinical Grade	
Patient Average Grade	Low-Intermediate Grade	High Grade
Low-Intermediate Grade	118	8
High Grade	45	117
	% Agreement	82
	kappa	0.64
	(95% CI)	(0.55-0.72)

[1]To assess agreement between two pathologists, patients were sampled from CBCS Phases 1, 2, and 3 for second pathology review.
[2]To assess agreement between image analysis and a pathologist, only samples with digital image data (CBCS3 only) were included.

Table A.2: Agreement between pathologists and between pathologists and image analysis in the test set for low-intermediate grade and high grade tumors

Unweighted						
Image Analysis	Negative	Positive	Sensitivity (%)	Specificity (%)	Accuracy (%)	Kappa (95% CI)
Overall						
ER negative	260	80	88	76	84	0.64 (0.59-0.69)
ER positive	83	572				
Low-Intermediate Grade						
ER negative	21	24	95	46	91	0.41 (0.28-0.55)
ER positive	25	467				
High Grade						
ER negative	239	46	69	80	77	0.49 (0.40-.57)
ER positive	58	104				
Grade-trained						
Image Analysis	Negative	Positive	Sensitivity (%)	Specificity (%)	Accuracy (%)	Kappa (95% CI)
Overall						
ER negative	246	104	84	72	80	0.55 (0.50-0.61)
ER positive	97	548				
Low-Intermediate Grade						
ER negative	28	69	86	61	84	0.31 (0.21-0.42)
ER positive	18	422				
High Grade						
ER negative	218	35	78	73	75	0.48 (0.44-0.56)
ER positive	79	125				

Table A.3: Impact of weighting by grade on accuracy, sensitivity, and specificity of ER status[1] in the test set.
[1]Numbers represent individual cores (n=995) from 288 patients with up to four cores per patient; H&E cores were excluded if missing IHC data (n=11)

among high grade tumors.

Risk of Recurrence. To examine the potential clinical relevance of using this image analysis technique we determined the sensitivity and specificity of image analysis and the ability to predict whether or not a tumor is classified as having a high vs. low-medium risk of recurrence score (ROR-PT) (Table A.4). ROR-PT is determined using a combination of tumor information including PAM50 subtype, tumor proliferation, and tumor size [Parker et al., 2009]. Overall, the accuracy of image analysis for ROR-PT was high at 76% (kappa 0.47). In grade-stratified analyses, accuracy for ROR-PT was higher among low-intermediate grade tumors (86%) than high grade tumors (67%).

	Intrinsic Subtype[2]					
	Basal-like	Non-Basal-like	Sensitivity (%)	Specificity (%)	Accuracy (%)	Kappa (95% CI)
Overall						
Basal-like	131	101	78	73	77	0.47 (0.32-0.54)
Non-Basal-like	48	368				
Low-Intermediate Grade						
Basal-like	11	41	86	73	85	0.27 (0.13-0.41)
Non-Basal-like	4	245				
High Grade						
Basal-like	120	60	67	73	70	0.40 (0.31-0.50)
Non-Basal-like	44	123				
	ROR-PT Status[2]					
	Low-Med	High	Sensitivity (%)	Specificity (%)	Accuracy (%)	Kappa (95% CI)
Overall						
Low-Med	342	40	79	74	76	0.47 (0.40-0.54)
High	118	148				
Low-Intermediate Grade						
Low-Med	245	16	47	90	86	0.32 (0.17-0.48)
High	26	14				
High Grade						
Low-Med	97	24	85	51	67	0.35 (0.26-0.44)
High	92	134				
	Histologic Subtype[3]					
	Ductal	Lobular	Sensitivity (%)	Specificity (%)	Accuracy (%)	Kappa (95% CI)
Overall						
Ductal	710	24	71	96	94	0.66 (0.57-0.74)
Lobular	28	58				
Low-Intermediate Grade						
Ductal	268	24	71	94	89	0.63 (0.53-0.73)
Lobular	23	58				
High Grade						
Ductal	442	0	N/A	99	99	N/A
Lobular	5	0				

Table A.4: Accuracy, sensitivity, and specificity of non-Basal-like intrinsic subtype, ROR-PT, and histologic subtype based on Image Analysis[1] in the test set.

[1]Numbers represent individual cores from patients where 1-4 cores were available. Cores were excluded if RNA data (n=358) was missing

[2]2180 patients with 648 cores for intrinsic subtype and ROR-PT

[3]3233 patients with 820 cores for histologic subtype

Histologic Subtype. In addition to using image analysis to predict tumor grade, we also tested this approach using histologic subtype, another visual feature of the tumor (Table A.4). Image analysis was able to predict a lobular compared to a ductal tumor with 94% accuracy (kappa 0.66). The accuracy was slightly lower when restricted to low grade tumors (89%) but was non-estimable among high grade tumors as there were no high grade lobular tumors in the test set.

Clinical Factors Associated with Classification Errors. To evaluate which clinical factors were associated with the accuracy of the image-based metrics, we evaluated predictors of accurate/inaccurate ER status calls (Table A.5) among patients in the test set (n=288). Considering age, race, grade, stage, lymph node status, ER status, Ki67 status, and mitotic tumor grade, no significant differences in accuracy of image-based ER assignment were observed. However, we found that image analysis tended to inaccurately predict ER status when tumors were Luminal B [OR, (95% CI); 4.42 (1.32-14.77)].

A.4 Discussion

In this study we used a deep learning approach to conduct image analysis on H&E-stained breast tumor tissue microarray samples from the population-based Carolina Breast Cancer Study, Phase 3 (2008-2013). First, we found that the agreement between image analysis and the pathologist-classified grade was only slightly lower than that observed for two study pathologists, and we obtained high agreement and kappa values. Second, we found that ER status, RNA-based molecular subtype (Basal-like vs. non-Basal like), and risk of recurrence score (ROR-PT) could be predicted with approximately 75-80% accuracy. Further, we found the image analysis accuracy to be 94% for ductal vs. lobular histologic subtype.

Grade. Previous literature based on comparing two pathologists shows that image assessment is subject to some disagreement [Longacre et al., 2006], particularly among the intermediate-grade tumors as we observed between the image analysis and pathologist classification in our study. Other groups have reported inter-rater kappa statistics of 0.6-0.7 for tumor grade [Longacre et al., 2006; Salles et al., 2008], in line with both our inter-pathologist agreement and

Variable	Inaccurate N (%)	Accurate N (%)	OR (95% CI)	Chi-squared p-value
Age				
≤50 years	23 (30.4)	110 (27.5)	Ref.	0.72
>50 years	20 (69.7)	135 (72.4)	0.71 (0.37-1.36)	
Race				
White	25 (84.3)	125 (77.6)	Ref.	0.22
Black	18 (15.7)	120 (22.4)	0.75 (0.39-1.45)	
Grade				
Low-Intermediate	19 (55.4)	143 (68.7)	Ref.	0.17
High	24 (44.6)	101 (31.3)	1.79 (0.93-3.44)	
Missing	0	1		
Stage				
I, II	39 (85.8)	220 (91.1)	Ref.	0.49
III, IV	4 (14.2)	25 (8.9)	0.90 (0.30-2.74)	
Node Status				
Negative	32 (73.8)	159 (68.2)	Ref.	0.53
Positive	11 (26.2)	86 (31.8)	0.64 (0.31-1.32)	
Tumor Size				
≤2cm	25 (60.0)	149 (68.6)	Ref.	0.38
>2cm	18 (40.0)	96 (31.4)	1.12 (0.58-2.16)	
IHC-based ER Status				
Negative	19 (37.8)	72 (20.2)	Ref.	0.07
Positive	24 (62.2)	173 (79.8)	0.53 (0.27-1.02)	
IHC-based Ki67 Status				
<10%	22 (56.1)	154 (67.8)	Ref.	0.24
≥10%	21 (43.9)	91 (32.2)	1.61 (0.84-3.10)	
Mitotic Grade				
1	14 (40.1)	116 (58.3)	Ref.	0.20
2	9 (20.9)	33 (12.4)	2.26 (0.90-5.68)	
3	20 (39.0)	95 (29.3)	1.74 (0.84-3.64)	
Missing	0	1		
Intrinsic Subtype				
Luminal A	5 (25.1)	69 (50.6)	Ref.	0.41
Luminal B	8 (26.5)	25 (20.0)	4.42 (1.32-14.77)	
Basal-like	9 (32.4)	40 (19.8)	3.10 (0.97-9.91)	
HER2	2 (12.4)	13 (4.8)	2.12 (0.37-12.14)	
Normal-like	2 (3.6)	7 (4.7)	3.94 (0.64-24.2)	
Missing	17	91		

Table A.5: Patient and tumor characteristics associated with inaccuracy of predicted ER status from the test set (n=288).

image analysis vs. pathologist agreement for grade. Elsewhere in the literature lower kappa values around 0.5 have been reported between pathologists for histologic grade [Boiesen et al., 2000]. In light of this inherent variability in image assessment, deep learning-based image analysis performed well at predicting tumor grade as low-intermediate vs. high using H&E images.

Molecular Markers. It is particularly promising that histologic subtype and molecular marker status could be predicted using image analysis. While we did perform grade-weighting within ER classification, there may be other image features of ER positive tumors that are not readily discernible and are driving the higher accuracy of ER positive images over ER negative. Agreement between true ER status (by IHC) and image analysis (kappa 0.64) was slightly lower than that observed for centralized pathology and SEER classifications for ER status (kappa 0.70) [Ma et al., 2009] and is similar to reports of agreement between different IHC antibodies for ER that show substantial agreement (kappa 0.6-0.8) [Prat et al., 2011]. Previous work with CBCS phase 1 samples found that agreement between medical records and staining of tissues was also similar (kappa of 0.62) [Carey et al., 2006]. Overall, the agreement between IHC-based ER status and image analysis predictions based on H&E-stained images are similar to estimates for comparing ER status classification in the literature. The high rate of agreement between pathologist-scored and image analysis-based histologic subtype was also compelling (kappa 0.64). Together these results suggest that some latent features indicative of underlying tumor biology are present in H&E images and can be identified through deep learning-based approaches.

Histologic Subtype. We observed a high accuracy of image analysis in predicting ductal versus lobular histologic subtype. The high accuracy may be due to the arrangement of epithelial and stromal cells characteristic of ductal and lobular tumors whereby lobular tumors are characterized by non-cohesive single file lines of epithelial cells infiltrating the stroma, and ductal tumors are characterized by sheets or nests of epithelial cells embedded in the surrounding stroma [Rosen, 2009; Makki, 2015]. We speculate that it may be that the high contrast staining between the epithelium and stromal components resulting from H&E immunohistochemistry

strengthens the ability of image analysis to predict this biologic feature of the tumor.

Intrinsic Subtype and Risk of Recurrence. With respect to intrinsic PAM50 subtype based solely upon gene expression values, previous studies have not evaluated image-based analysis for predicting intrinsic subtype or the risk of recurrence using a score-based method, ROR-PT [Parker et al., 2009]. A few previous studies have evaluated the clinical record or a central immunohistochemistry laboratory vs. RNA-based subtyping for Basal-like vs. non-Basal-like. Even considering two molecular comparisons, agreements do not exceed 90% - that is, Allott et al. (2015) found approximately 90% agreement between Basal-like status for IHC-based vs. RNA-based assessment and 77% agreement for classification of Luminal A subtype [Allott et al., 2015]. Our estimates are similar suggesting that image analysis, even without the use of special IHC stains, could be a viable option for classification of molecular breast tumor subtype and ROR-PT from H&E-stained images.

Limitations. As with other studies, our work should be viewed in light of some limitations. Our sample size was limited in our testing set to 288 patients, but this resulted in nearly 1,000 TMA cores available for use in our image analysis. Using a larger set of samples with data on RNA-based subtype to balance training for each predictor could be useful. For example, the fact that Luminal B patients had a higher error rate might suggest there are some features of Luminal B breast cancers that are distinct and image-detectable and a larger sample size would be helpful in identifying these. Deep learning may be utilizing these features, but, in our small sample set, we are unable to tune our data to specifically identify those features or to clarify what they are in intuitive language. In addition, the use of binary classification systems for training our digital algorithms (i.e., Basal-like vs. non-Basal-like) does not allow us to differentiate among all five RNA-based intrinsic subtypes. Currently, U.S.-based genomic tests provide continuous risk scores but also suggest relevant cut points that in essence make these assays almost a binary classification; therefore, binary classification may have some utility in the current clinical context. However, future work should extend these approaches to multi-class classification. Furthermore, improved results may be obtained by fine-tuning the Convolutional Neural Network for breast cancer H&E image classification.

Clinical Value. Image-based risk prediction has potential clinical value. Gene expression data on tumor tissue samples is not uniformly available for all patients and is costly to obtain in both a clinical and epidemiologic setting. These results suggest that tumor histology and molecular subtype along with the risk of recurrence (ROR-PT) can be predicted from H&E images alone in a high-throughput, objective, and accurate manner. These results could be used to identify patients who would benefit from further genomic testing. Furthermore, even ER testing is not routinely performed in countries with limited laboratory testing resources, and predicting ER status by morphologic features may have utility for guiding endocrine therapy in low-resource settings.

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283.

Agusti, A., Calverley, P., Celli, B., Coxson, H. O., Edwards, L. D., Lomas, D. A., MacNee, W., Miller, B. E., Rennard, S., Silverman, E. K., et al. (2010). Characterisation of ocpd heterogeneity in the eclipse cohort. *Respir Res*, 11(1):122.

Alizadeh, A. A., Aranda, V., Bardelli, A., Blanpain, C., Bock, C., Borowski, C., Caldas, C., Califano, A., Doherty, M., Elsner, M., Esteller, M., Fitzgerald, R., Korbel, J. O., Lichter, P., Mason, C. E., Navin, N., Pe'er, D., Polyak, K., Roberts, C. W. M., Siu, L., Snyder, A., Stower, H., Swanton, C., Verhaak, R. G. W., Zenklusen, J. C., Zuber, J., and Zucman-Rossi, J. (2015). Toward understanding and exploiting tumor heterogeneity. *Nature Medicine*, 21(8):846–853.

Allott, E. H., Cohen, S. M., Geradts, J., Sun, X., Khoury, T., Zirpoli, G. R., Miller, C. R., Hwang, H., Thorne, L. B., Connor, S. O., Tse, C.-k., Bell, M. B., Hu, Z., Li, Y., Kirk, E. L., Bethea, T. N., Perou, C. M., Palmer, J. R., Ambrosone, C. B., Olshan, A. F., and Troester, M. A. (2015). Performance of three biomarker immunohistochemistry for intrinsic breast cancer subtyping in the AMBER consortium. *Cancer Epidemiology Biomarkers & Prevention*, pages 1–28.

Andrew, G., Arora, R., Bilmes, J., and Livescu, K. (2013). Deep Canonical Correlation Analysis. In *Proc. ICML*, volume 28.

Andrews, S., Tsochantaridis, I., and Hofmann, T. (2002). Support vector machines for multiple-instance learning. In *Advances in neural information processing systems*, pages 561–568.

Arevalo, J., Cruz-Roa, A., Arias, V., Romero, E., and González, F. A. (2015). An unsupervised feature learning framework for basal cell carcinoma image analysis. *Artificial Intelligence in Medicine*, 64(2):131–145.

Azizpour, H., Razavian, A. S., Sullivan, J., Maki, A., and Carlsson, S. (2014). Factors of Transferability for a Generic ConvNet Representation. *arXiv preprint: 1406.5774*.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Bahrampour, S., Nasrabadi, N. M., Ray, A., and Jenkins, W. K. (2015). Multimodal Task-Driven Dictionary Learning for Image Classification. *arXiv preprint: 1502.01094*.

Basavanhally, A., Yu, E., Xu, J., Ganesan, S., Feldman, M., Tomaszewski, J., and Madabhushi, A. (2011). Incorporating domain knowledge for tubule detection in breast histopathology using O'Callaghan neighborhoods. *Proc. SPIE*, 7963(796310).

Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359.

Beck, A. H., Sangoi, A. R., Leung, S., Marinelli, R. J., Nielsen, T. O., van de Vijver, M. J., West, R. B., van de Rijn, M., and Koller, D. (2011). Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Science Translational Medicine*, 3(108):108–113.

Bentaieb, A. and Hamarneh, G. (2018). Adversarial stain transfer for histopathology image analysis. *IEEE transactions on medical imaging*, 37(3):792–802.

BenTaieb, A. and Hamarneh, G. (2018). Predicting cancer with a recurrent visual attention model for histopathology images. In *Proc. MICCAI*. Springer.

Bhatt, G., Jha, P., and Raman, B. (2017). Common Representation Learning Using Step-based Correlation Multi-Modal CNN. *arXiv preprint: 1711.00003*.

Bie, T. D., Cristianini, N., and Rosipal, R. (2005). Eigenproblems in pattern recognition. In *Handbook of Geometric Computing*, pages 129–167. Springer Berlin Heidelberg.

Bilenko, N. Y. and Gallant, J. L. (2016). Pyrcca: regularized kernel canonical correlation analysis in Python and its applications to neuroimaging. *Frontiers in Neuroinformatics*, 10.

Boiesen, P., Bendahl, P.-O., Anagnostaki, L., Domanski, H., Holm, E., Idvall, I., Johansson, S., Ljungberg, O., and Ringberg, A. (2000). Histologic grading in breast cancer: reproducibility between seven pathologic departments. *Acta oncologica*, 39(1):41–45.

Broadhurst, R. E. (2008). *Compact appearance in object populations using quantile function based distribution families*. PhD thesis, The University of North Carolina at Chapel Hill.

Broekaert, S. M. C., Roy, R., Okamoto, I., van den Oord, J., Bauer, J., Garbe, C., Barnhill, R. L., Busam, K. J., Cochran, A. J., Cook, M. G., Elder, D. E., McCarthy, S. W., Mihm, M. C., Schadendorf, D., Scolyer, R. a., Spatz, A., and Bastian, B. C. (2010). Genetic and morphologic features for melanoma classification. *Pigment Cell Melanoma Research*, 23(6):763–70.

Carbonneau, M.-A., Cheplygina, V., Granger, E., and Gagnon, G. (2017). Multiple Instance Learning: A Survey on Problems Characteristics and Applications. *arXiv preprint: 1612.03365*.

Carey, L. A., Perou, C. M., Livasy, C. A., Dressler, L. G., Cowan, D., Conway, K., Karaca, G., Troester, M. a., Tse, C. K., Edmiston, S., Deming, S. L., Geradts, J., Cheang, M. C. U., Nielsen, T. O., Moorman, P. G., Earp, H. S., and Millikan, R. C. (2006). Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. *JAMA : the journal of the American Medical Association*, 295(21):2492–502.

Carlson, J. J. and Roth, J. A. (2013). The impact of the Oncotype Dx breast cancer assay in clinical practice: a systematic review and meta-analysis. *Breast Cancer Res Treat*, 141(1):13–22.

Caron, M., Bojanowski, P., Joulin, A., and Douze, M. (2018). Deep clustering for unsupervised learning of visual features. In *Proc. ECCV*.

Cawley, G. C. and Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(Jul):2079–2107.

Cha, M., Gwon, Y., and Kung, H. T. (2015). Multimodal sparse representation learning and applications. *arXiv preprint: 1511.06238*.

Chandar, S., Khapra, M. M., Larochelle, H., and Ravindran, B. (2016). Correlational Neural Networks. *Neural Computation*, 28(2):257–285.

Chang, H., Fontenay, G. V., Han, J., Cong, G., Baehner, F. L., Gray, J. W., Spellman, P. T., and Parvin, B. (2011). Morphometic analysis of TCGA glioblastoma multiforme. *BMC Bioinformatics*, 12(1):484.

Chang, H., Nayak, N., Spellman, P. T., and Parvin, B. (2013). Characterization of Tissue Histopathology via Predictive Sparse Decomposition and Spatial Pyramid Matching. In *Proc. MICCAI*.

Chang, H., Zhou, Y., Borowsky, A., Barner, K., Spellman, P., and Parvin, B. (2015). Stacked Predictive Sparse Decomposition for Classification of Histology Sections. *IJCV*, 113(1):3–18.

Chang, X., Xiang, T., and Hospedales, T. M. (2018). Scalable and Effective Deep CCA via Soft Decorrelation. In *Proc. CVPR*.

Chen, Y., Bi, J., and Wang, J. Z. (2006). MILES: multiple-instance learning via embedded instance selection. *IEEE PAMI*, 28(12):1931–47.

Chen, Y. and Wang, J. Z. (2004). Image Categorization by Learning and Reasoning with Regions. *The Journal of Machine Learning Research*, 5:913–939.

Cheplygina, V., Sørensen, L., Tax, D. M. J., Bruijne, M. D., and Loog, M. (2015). Label Stability in Multiple Instance Learning. In *Proc. MICCAI*.

Chu, W. and Keerthi, S. S. (2007). Support vector ordinal regression. *Neural computation*, 19(3):792–815.

Cireşan, D. C., Giusti, A., Gambardella, L. M., and Schmidhuber, J. (2013). Mitosis detection in breast cancer histology images with deep neural networks. In *Proc. MICCAI*.

Coates, A., Lee, H., and Ng, A. (2010). An analysis of single-layer networks in unsupervised feature learning. In *Proc. AISTATS*.

Coates, A. and Ng, A. (2011). The importance of encoding versus training with sparse coding and vector quantization. In *Proc. ICML*.

Codella, N., Cai, J., Abedini, M., and Garnavi, R. (2015). Deep Learning, Sparse Coding, and SVM for Melanoma Recognition in Dermoscopy Images. In *Proc. MICCAI Workshop on Machine Learning in Medical Imaging*.

Cogswell, M., Ahmed, F., Girshick, R., Zitnick, L., and Batra, D. (2016). Reducing Overfitting in Deep Networks by Decorrelating Representations. *Proc. ICLR*.

Cooper, L. A. D., Kong, J., Gutman, D. A., Wang, F., Gao, J., Appin, C., Cholleti, S., Pan, T., Sharma, A., Scarpace, L., Mikkelsen, T., Kurc, T., Moreno, C. S., Brat, D. J., and Saltz, J. H. (2012). Integrated morphologic analysis for the identification and characterization of disease subtypes. *Journal of the American Medical Informatics Association*, 19(2):317–23.

Couture, H. D., Marron, J. S., Perou, C. M., Troester, M. A., and Niethammer, M. (2018a). Multiple instance learning for heterogeneous images: Training a CNN for histopathology. In *Proc. MICCAI*.

Couture, H. D., Marron, J. S., Thomas, N. E., Perou, C. M., and Niethammer, M. (2015). Hierarchical task-driven feature learning for tumor histology. In *Proc. ISBI*.

Couture, H. D., Williams, L., Geradts, J., Nyante, S., Butler, E., Marron, J., Perou, C., Troester, M., and Niethammer, M. (2018b). Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. *npj Breast Cancer*.

Cox, D. R. (1992). Regression models and life-tables. In *Breakthroughs in statistics*, pages 527–541. Springer.

Cruz-Roa, A., Arévalo, J., Judkins, A., Madabhushi, A., and González, F. (2015). A method for medulloblastoma tumor differentiation based on convolutional neural networks and transfer learning. In *Proc. International Symposium on Medical Information Processing and Analysis*.

Cruz-Roa, A., Basavanhally, A., González, F., Gilmore, H., Feldman, M., Ganesan, S., Shih, N., Tomaszewski, J., and Madabhushi, A. (2014). Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. *Proc. SPIE*, 9041(216).

Cruz-Roa, A., Caicedo, J. C., and González, F. a. (2011). Visual pattern mining in histology image collections using bag of features. *Artificial Intelligence in Medicine*, 52(2):91–106.

Cruz-Roa, A., Gilmore, H., Basavanhally, A., Feldman, M., Ganesan, S., Shih, N. N., Tomaszewski, J., González, F. A., and Madabhushi, A. (2017). Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent. *Scientific Reports*, 7:46450.

Cruz-Roa, A. A., Ovalle, J. E. A., Madabhushi, A., and Gonzalez, F. A. O. (2013). A Deep Learning Architecture for Image Representation, Visual Interpretability and Automated Basal-Cell Carcinoma Cancer Detection. In *Proc. MICCAI*.

Csurka, G., Dance, C., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proc. CVPR*, volume 1, pages 886–893.

DeVries, P. M., Viégas, F., Wattenberg, M., and Meade, B. J. (2018). Deep learning of aftershock patterns following large earthquakes. *Nature*, 560(7720):632.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. (2014). DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In *Proc. ICML*.

Dorfer, M., Kelz, R., and Widmer, G. (2016a). Deep linear discriminant analysis. In *Proc. ICLR*.

Dorfer, M., Schlüter, J., Vall, A., Korzeniowski, F., and Widmer, G. (2018). End-to-end cross-modality retrieval with CCA projections and pairwise ranking loss. *International Journal of Multimedia Information Retrieval*, 7(2):117–128.

Dorfer, M., Widmer, G., and At, G. W. (2016b). Towards Deep and Discriminative Canonical Correlation Analysis. In *Proc. ICML Workshop on Multi-view Representaiton Learning*.

Duan, K., Zhang, H., and Wang, J. J. Y. (2016). Joint learning of cross-modal classifier and factor analysis for multimedia data classification. *Neural Computing and Applications*, 27(2):459–468.

Dunnwald, L. K., Rossing, M. A., and Li, C. I. (2007). Hormone receptor status, tumor characteristics, and prognosis: a prospective cohort of breast cancer patients. *Breast cancer research : BCR*, 9(1):R6.

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

Gotlieb, C. and Kreyszig, H. (1990). Texture descriptors based on co-occurrence matrices. *Computer Vision, Graphics and Image Processing*, 51:76–80.

Han, J., Chang, H., Loss, L., Zhang, K., Baehner, F. L., Gray, J. W., Spellman, P., and Parvin, B. (2011). Comparison of sparse coding and kernel methods for histopathological classification of gliobastoma multiforme. In *Proc. ISBI*, pages 711–714.

Hardoon, D. R. and Shawe-Taylor, J. (2011). Sparse canonical correlation analysis. *Machine Learning*, 83(3):331–353.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proc. CVPR*.

Hiley, C. T. and Swanton, C. (2014). Spatial and temporal cancer evolution: causes and consequences of tumour diversity. *Clinical Medicine*, 14(Suppl_6):s33–s37.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.

Hou, L., Agarwal, A., Samaras, D., Kurc, T. M., Gupta, R. R., and Saltz, J. H. (2017). Unsupervised histopathology image synthesis. *arXiv preprint: 1712.05021*.

Hou, L., Samaras, D., Kurc, T. M., Gao, Y., Davis, J. E., Saltz, J. H., Hospital, S. B., and Hospital, S. B. (2016). Patch-based Convolutional Neural Network for Whole Slide Tissue Image Classification. In *Proc. CVPR*.

Huang, C., Li, Y., Change Loy, C., and Tang, X. (2016). Learning deep representation for imbalanced classification. In *Pro. CVPR*.

Huang, C., Zhang, A., and Xiao, G. (2018a). Deep Integrative Analysis for Survival Prediction. In *Proc. Pacific Symposium on Biocomputing*.

Huang, L., Yang, D., Lang, B., and Deng, J. (2018b). Decorrelated Batch Normalization. In *Proc. CVPR*.

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Networks*, 13(4):411–430.

Ioffe, S. and Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proc. ICML*.

Janowczyk, A. and Anant, M. (2016). Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of Pathology Informatics*.

Jia, Z., Huang, X., Chang, E. I.-C., and Xu, Y. (2017). Constrained Deep Weak Supervision for Histopathology Image Segmentation. *arXiv preprint: 1701.00794*.

Jiang, Z., Lin, Z., and Davis, L. S. (2013). Label Consistent K-SVD: Learning a Discriminative Dictionary for Recognition. *IEEE PAMI*, 35(11):2651–64.

Jiao, C. and Zare, A. (2015). Multiple Instance Dictionary Learning using Functions of Multiple Instances. *arXiv preprint: 1511.02825*.

Julesz, B. (1981). Textons, the elements of texture perception, and their interactions. *Nature*, 290:91–97.

Kan, M., Shan, S., Zhang, H., Lao, S., and Chen, X. (2015). Multi-view Discriminant Analysis. *IEEE PAMI*.

Kandemir, M. and Hamprecht, F. A. F. (2014). Computer-aided diagnosis from weak supervision: A benchmarking study. *Computerized Medical Imaging and Graphics*.

Katzman, J., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. (2016). Deep Survival: A Deep Cox Proportional Hazards Network. *arxiv preprint: 1606.00931*.

Kessy, A., Lewin, A., and Strimmer, K. (2015). Optimal whitening and decorrelation. *arXiv preprint: 1512.00809*.

Khan, A. M., Sirinukunwattana, K., and Rajpoot, N. (2015). A Global Covariance Descriptor for Nuclear Atypia Scoring in Breast Histopathology Images. *IEEE Journal of Biomedical and Health Informatics*, 19(5):1637–1647.

Khan, S. H., Hayat, M., Bennamoun, M., Sohel, F. A., and Togneri, R. (2017). Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*.

Kong, B., Sun, S., Wang, X., Song, Q., and Zhang, S. (2018). Invasive cancer detection utilizing compressed convolutional neural network and transfer learning. In *Proc. MICCAI*. Springer.

Kraus, O. Z., Ba, J. L., and Frey, B. J. (2016). Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics*, 32(12):i52–i59.

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232.

Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1106–1114.

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

Lai, P. L. and Fyfe, C. (2000). Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 10(5):365–377.

Lambert, J.-C. and Amouyel, P. (2007). Genetic heterogeneity of alzheimer's disease: complexity and advances. *Psychoneuroendocrinology*, 32:S62–S70.

Le, Q. V., Han, J., Gray, J. W., Spellman, P. T., Borowsky, A., and Parvin, B. (2012a). Learning invariant features of tumor signatures. In *Proc. ISBI*, pages 302–305.

Le, Q. V., Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G. S., Dean, J., and Ng, A. Y. (2012b). Building High-Level Features using Large Scale Unsupervised Learning. In *Proc. ICML*.

LeCun, Y. (1998). The mnist database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Lee, G., Singanamalli, A., Wang, H., Feldman, M. D., Master, S. R., Shih, N. N. C., Spangler, E., Rebbeck, T., Tomaszewski, J. E., and Madabhushi, A. (2015). Supervised multi-view canonical correlation analysis (sMVCCA): integrating histologic and proteomic features for predicting recurrent prostate cancer. *IEEE transactions on medical imaging*, 34(1):284–97.

Lepistö, L., Kunttu, I., Autio, J., and Visa, A. (2003). Classification method for colored natural textures using gabor filtering. In *Proceedings of 12th International Conference on Image Analysis and Processing*, pages 397–401.

Li, D., Dimitrova, N., Li, M., and Sethi, I. K. (2003). Multimedia content processing through cross-modal association. In *Proc. ACM International Conference on Multimedia*.

Li, L. and Lin, H.-T. (2007). Ordinal regression by extended binary classification. In *Advances in neural information processing systems*, pages 865–872.

Li, W. and Vasconcelos, N. (2015). Multiple instance learning for soft bags via top instances. In *Proc. CVPR*.

Li, W., Zhang, J., and Mckenna, S. J. (2015). Multiple Instance Cancer Detection by Boosting Regularised Trees. In *Proc. MICCAI*.

Li, Y., Wang, J., Ye, J., and Reddy, C. K. (2016). A Multi-Task Learning Formulation for Survival Analysis. In *Proc. International Conference on Knowledge Discovery and Data Mining*. ACM Press.

Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2013). Joint and Individual Variation Explained (JIVE) for Integrated Analysis of Multiple Data Types. *The Annals of Applied Statistics*, 7(1):523–542.

Longacre, T. A., Ennis, M., Quenneville, L. A., Bane, A. L., Bleiweiss, I. J., Carter, B. A., Catelano, E., Hendrickson, M. R., Hibshoosh, H., Layfield, L. J., et al. (2006). Interobserver agreement and reproducibility in classification of invasive breast carcinoma: an nci breast cancer family registry study. *Modern pathology*, 19(2):195.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

Lu, L., Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Member, S., Xu, Z., Nogues, I., Yao, J., Mollura, D., and Summers, R. M. (2016). Deep Convolutional Neural Networks for Computer-Aided Detection : CNN Architectures , Dataset Characteristics and Transfer Learning Deep Convolutional Neural Networks for Computer-Aided Detection : CNN Architectures , Dataset Characteristics and Transfer. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298.

Ma, H., Wang, Y., Sullivan-Halley, J., Weiss, L., Burkman, R., Simon, M., Malone, K., Strom, B., Ursin, G., Marchbanks, P., McDonald, J., Spirta, R., Press, M., and Bernstein, L. (2009). Breast Cancer Receptor Status: Do Results from a Centralized Pathology Laboratory Agree with SEER Registry Reports? *Cancer Epidemiol Biomarkers Prev*, 18(8):2214–2220.

Mairal, J., Bach, F., and Ponce, J. (2012). Task-driven dictionary learning. *IEEE PAMI*, 34(4):791–804.

Mairal, J., Bach, F., and Ponce, J. (2014). Sparse Modeling for Image and Vision Processing. In *Foundations and Trends in Computer Graphics and Vision*, volume 8, pages 49–50.

Mairal, J., Bach, F., Ponce, J., and Sapiro, G. (2009). Online dictionary learning for sparse coding. In *Proc. ICML*.

Makki, J. (2015). Diversity of Breast Carcinoma: Histological Subtypes and Clinical Relevance. *Clinical Medicine Insights: Pathology*, 8:23–31.

Mannino, D. M. (2002). Copd: epidemiology, prevalence, morbidity and mortality, and disease heterogeneity. *CHEST Journal*, 121(5_suppl):121S–126S.

Masters, D. and Luschi, C. (2018). Revisiting Small Batch Training for Deep Neural Networks. *arxiv preprint: 1804.07612*.

McGranahan, N. and Swanton, C. (2015). Biological and Therapeutic Impact of Intratumor Heterogeneity in Cancer Evolution. *Cancer Cell*, 27(1):15–26.

Miedema, J., Marron, J. S., Niethammer, M., Borland, D., Woosley, J., Coposky, J., Wei, S., Reisner, H., and Thomas, N. E. (2012). Image and statistical analysis of melanocytic histology. *Histopathology*, 61(3):436–44.

Natrajan, R., Sailem, H., Mardakheh, F. K., Garcia, M. A., Tape, C. J., Dowsett, M., Bakal, C., and Yuan, Y. (2016). Microenvironmental heterogeneity parallels breast cancer progression: a histology–genomic integration analysis. *PLoS medicine*, 13(2):e1001961.

Nayak, N., Chang, H., Borowsky, A., Spellman, P., and Parvin, B. (2013). Classification of tumor histopathology via sparse feature learning. In *Proc. ISBI*, pages 410–413.

Network, C. G. A. et al. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61.

Niethammer, M., Borland, D., Marron, J., Woolsey, J., and Thomas, N. (2010). Appearance normalization of histology slides. In *Proc. MICCAI, International Workshop on Machine Learning in Medical Imaging*.

Odena, A. (2016). Semi-supervised learning with generative adversarial networks. *arXiv preprint: 1606.01583*.

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. (2014). Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks. In *Proc. CVPR*.

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F. L., Walker, M. G., Watson, D., Park, T., et al. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine*, 351(27):2817–2826.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*, 27(8):1160–1167.

Prat, A., Ellis, M. J., and Perou, C. M. (2011). Practical implications of gene-expression-based assays for breast oncologists. *Nature reviews. Clinical oncology*, 9(1):48–57.

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

Ranzato, M. A. and Szummer, M. (2008). Semi-supervised learning of compact document representations with deep networks. In *Proc. ICML*, pages 792–799.

Ray, B., Henaff, M., Ma, S., Efstathiadis, E., Peskin, E. R., Picone, M., Poli, T., Aliferis, C. F., and Statnikov, A. (2014a). Information content and analysis methods for multi-modal high-throughput biomedical data. *Scientific reports*, 4:4411.

Ray, P., Zheng, L., Lucas, J., and Carin, L. (2014b). Bayesian joint analysis of heterogeneous genomics data. *Bioinformatics*, 30(10):1370–6.

Razavian, A. S., Azizpour, H., Sullivan, J., and Carlsson, S. (2014). CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In *Proc. CVPR*, pages 512–519. IEEE.

Rosen, P. P. (2009). *Rosen's Breast Pathology*. Lippincott Williams & Wilkins, Phildelphia, third edition.

Salles, M., Sanches, F., and Perez AA, G. (2008). Importance of a second opinion in breast surgical pathology and therapeutic implications. *Revista brasileira de ginecologia e obstetricia : revista da Federacao Brasileira das Sociedades de Ginecologia e Obstetricia*, 30(12):602–608.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. (2014). OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. In *Proc. International Conference on Learning Representations*.

Shaban, M. T., Baur, C., Navab, N., and Albarqouni, S. (2018). Staingan: Stain style transfer for digital histological images. *arXiv preprint: 1804.01601*.

Shekhar, S., Patel, V. M., Nasrabadi, N. M., and Chellappa, R. (2014). Joint sparse representation for robust multimodal biometrics recognition. *IEEE PAMI*, 36(1):113–26.

Shouno, H., Suzuki, S., and Kido, S. (2015). A Transfer Learning Method with Deep Convolutional Neural Network for Diffuse Lung Disease Classification. In *Advances in neural information processing systems*.

Shrivastava, A., Patel, V. M., Pillai, J. K., and Chellappa, R. (2015). Generalized Dictionaries for Multiple Instance Learning. *IJCV*, 114(2-3):288–305.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint: 1312.6034*.

Simonyan, K. and Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proc. International Conference on Learning Representations*.

Singanamalli, A., Wang, H., Lee, G., Shih, N., Rosen, M., Master, S., Tomaszewski, J., Feldman, M., and Madabhushi, A. (2014). Supervised multi-view canonical correlation analysis: fused multimodal prediction of disease diagnosis and prognosis. In *Proc. SPIE Medical Imaging*, page 903805.

Song, X., Jiao, L., Yang, S., Zhang, X., and Shang, F. (2013). Sparse coding and classifier ensemble based multi-instance learning for image categorization. *Signal Processing*, 93(1):1–11.

Spanhol, F. A., Oliveira, L. S., Cavalin, P. R., Petitjean, C., and Heutte, L. (2017). Deep features for breast cancer histopathological image classification. In *Proc. International Conference on Systems, Man, and Cybernetics*, pages 1868–1873. IEEE.

Spanhol, F. A., Oliveira, L. S., Petitjean, C., and Heutte, L. (2016a). A Dataset for Breast Cancer Histopathological Image Classification. *IEEE Transactions on Biomedical Engineering*, 63(7):1455–1462.

Spanhol, F. A., Oliveira, L. S., Petitjean, C., and Heutte, L. (2016b). Breast Cancer Histopathological Image Classification using Convolutional Neural Networks. In *International Joint Conference on Neural Networks*.

Sparano, J. A. and Paik, S. (2008). Development of the 21-gene assay and its application in clinical practice and clinical trials. *Journal of Clinical Oncology*, 26(5):721–728.

Springenberg, J. T. (2015). Unsupervised and semi-supervised learning with categorical generative adversarial networks. *arXiv preprint: 1511.06390*.

Sun, M., Han, T. X., Liu, M.-C., and Khodayari-Rostamabad, A. (2016). Multiple Instance Learning Convolutional Neural Networks for Object Recognition. In *Proc. ICPR*.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, page 12.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going Deeper With Convolutions. In *Proc. CVPR*.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint: 1312.6199*.

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., and Liang, J. (2016). Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? *IEEE Transactions on Medical Imaging*, 35(5):1299–1312.

Tavassoli, F. A. and Devilee, P. (2003). Tumors of the Breast. In *Pathology & Genetics: Tumours of the Breast and Female Genital Organs*, chapter 1. Iarc.

Troester, M., Sun, X., Allott, E. H., Geradts, J., Cohen, S. M., Tse, C. K., Kirk, E. L., Thorne, L. B., Matthews, M., Li, Y., Hu, Z., Robinson, W. R., Hoadley, K. A., Olopade, O. I., Reeder-Hayes, K. E., Earp, H. S., Olshan, A. F., Carey, L., and Perou, C. M. (2018). Racial differences in PAM50 subtypes in the Carolina Breast Cancer Study. *Journal of the National Cancer Institute*.

Tuceryan, M. and Jain, A. (1998). *Texture Analysis*, pages 207–248. World Scientific Publishing Co., 2nd edition.

Van Der Maaten, L. and Hinton, G. (2008). Visualizing high-dimensional data using t-sne. journal of machine learning research. *Journal of Machine Learning Research*, 9:26.

Vanwinckelen, G., Tragante do O, V., Fierens, D., and Blockeel, H. (2016). Instance-level accuracy versus bag-level accuracy in multi-instance learning. *Data Mining and Knowledge Discovery*, 30(2):313–341.

Varma, M. and Zisserman, A. (2005). A statistical approach to texture classification from single images. *International Journal of Computer Vision: Special Issue on Texture Analysis and Synthesis*, 62(1-2):61–81.

Varma, M. and Zisserman, A. (2007). A statistical approach to material classification using image patch exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Varol, E., Sotiras, A., and Davatzikos, C. (2015). Disentangling Disease Heterogeneity with Max-Margin Multiple Hyperplane Classifier. In *Proc. MICCAI*.

Veta, M., van Diest, P. J., Willems, S. M., Wang, H., Madabhushi, A., Cruz-Roa, A., Gonzalez, F., Larsen, A. B. L., Vestergaard, J. S., Dahl, A. B., Cireşan, D. C., Schmidhuber, J., Giusti, A., Gambardella, L. M., Tek, F. B., Walter, T., Wang, C.-W., Kondo, S., Matuszewski, B. J., Precioso, F., Snell, V., Kittler, J., de Campos, T. E., Khan, A. M., Rajpoot, N. M., Arkoumani, E., Lacle, M. M., Viergever, M. A., and Pluim, J. P. W. (2015). Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Medical image analysis*, 20(1):237–48.

Vu, T. H., Mousavi, H. S., Monga, V., Rao, U. A., and Rao, G. (2015). DFDL: Discriminative Feature-oriented Dictionary Learning for Histopathological Image Classification. In *Prob. ISBI*.

Wang, D., Khosla, A., Gargeya, R., Irshad, H., and Beck, A. H. (2016a). Deep Learning for Identifying Metastatic Breast Cancer. *arXiv preprint: 1606.05718*.

Wang, L., Li, Y., Zhou, J., Zhu, D., and Ye, J. (2017). Multi-task Survival Analysis. In *Proc. International Conference on Data Mining*, pages 485–494. IEEE.

Wang, W., Arora, R., Livescu, K., and Bilmes, J. (2015a). On deep multi-view representation learning. In *Proc. ICML*.

Wang, W., Arora, R., Livescu, K., and Srebro, N. (2016b). Stochastic optimization for deep CCA via nonlinear orthogonal iterations. In *Proc. Allerton Conference on Communication, Control, and Computing*.

Wang, X., Wang, B., Bai, X., Liu, W., and Tu, Z. (2013). Max-Margin Multiple-Instance Dictionary Learning. In *Proc. ICML*.

Wang, X., Yan, Y., Tang, P., Bai, X., and Liu, W. (2018). Revisiting Multiple Instance Neural Networks. *Pattern Recognition*, 74:15–24.

Wang, X., Zhu, Z., Yao, C., and Bai, X. (2015b). Relaxed Multiple-Instance SVM with Application to Object Discovery. *arXiv preprint: 1510.01027*.

Wegelin, J. (2000). A survey of Partial Least Squares (PLS) methods, with emphasis on the two-block case. Technical report, Department of Statistics, University of Washington, Seattle.

Xu, J., Luo, X., Wang, G., Gilmore, H., and Madabhushi, A. (2016). A Deep Convolutional Neural Network for Segmenting and Classifying Epithelial and Stromal Regions in Histopathological Images. *Neurocomputing*, 191:214–223.

Xu, X., Shimada, A., Taniguchi, R.-i., and He, L. (2015). Coupled dictionary learning and feature mapping for cross-modal retrieval. In *Proc. International Conference on Multimedia and Expo*.

Xu, Y., Mo, T., Feng, Q., Zhong, P., Lai, M., and Chang, E. I.-C. (2014a). Deep learning of feature representation with multiple instance learning for medical image analysis. In *Proc. International Conference on Acoustics, Speech and Signal Processing*.

Xu, Y., Zhu, J. Y., Chang, E. I. C., Lai, M., and Tu, Z. (2014b). Weakly supervised histopathology cancer image segmentation and classification. *Medical Image Analysis*, 18(3):591–604.

Yang, Z. and Michailidis, G. (2015). A Non-negative Matrix Factorization Method for Detecting Modules in Heterogeneous Omics Multi-modal Data. *Bioinformatics*.

Yao, J., Zhu, X., Zhu, F., and Huang, J. (2017). Deep correlational learning for survival prediction from multi-modality data. In *Proc. MICCAI*.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems*.

Young, B., Woodford, P., and O'Dowd, G. (2013). *Wheater's Functional Histology E-Book: A Text and Colour Atlas*. Elsevier Health Sciences.

Yuan, Y., Failmezger, H., Rueda, O. M., Ali, H. R., Gräf, S., Chin, S.-F., Schwarz, R. F., Curtis, C., Dunning, M. J., Bardwell, H., Johnson, N., Doyle, S., Turashvili, G., Provenzano, E., Aparicio, S., Caldas, C., and Markowetz, F. (2012). Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Science translational medicine*, 4(157).

Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pages 694–699.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Proc. ECCV*.

Zhang, W., Li, R., Zeng, T., Sun, Q., Kumar, S., Ye, J., and Ji, S. (2015). Deep Model Based Transfer and Multi-Task Learning for Biological Image Analysis. In *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Zhou, B., Bau, D., Oliva, A., and Torralba, A. (2017). Interpreting Deep Visual Representations via Network Dissection. *arXiv preprint: 1711.05611*.

Zhou, G., Cichocki, A., Zhang, Y., and Mandic, D. P. (2015). Group Component Analysis for Multiblock Data: Common and Individual Feature Extraction. *IEEE transactions on neural networks and learning systems*.

Zhou, Y., Chang, H., Barner, K., Spellman, P., and Parvin, B. (2014). Classification of histology sections via multispectral convolutional sparse coding. In *Proc. CVPR*.

Zhou, Z.-H., Sun, Y.-Y., and Li, Y.-F. (2009). Multi-instance learning by treating instances as non-I.I.D. samples. In *Proc. ICML*.