ARTIFICIALLY INTELLIGENT PATHOLOGY

By

Geoffrey Fredrick Schau

A DISSERTATION

Presented to
The Department of Biomedical Engineering
School of Medicine
Oregon Health & Science University

In partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

August 2020

Certificate of Approval

This is to certify that the PhD dissertation of

Geoffrey Fredrick Schau

has been approved by

Young Hwan Chang, Mentor & Advisor

Laura M. Heiser, Committee Chair

Joe W. Gray, Committee Member

Christopher Corless, Committee Member

Andrew Adey, Committee Member

Guillaume Thibault, Committee Member

Xubo Song, External Committee Member

To Mom and Dad

Contents

1 Introduction **1**

1.1 Spatial Systems Biology of Cancer 2

1.1.1 Mechanisms by Which Cancer Arises 3

1.1.2 Metastasis 6

1.1.3 Imaging Science 6

1.2 Learning Systems Architecture 11

1.2.1 Artificial Neural Networks 11

1.2.2 Convolutional Neural Networks 13

1.2.3 Updating Learning Models 18

1.3 Digital Pathology & Artificial Intelligence 25

1.4 Challenges and Opportunities 30

1.5 Summary 33

1.5.1 Contributions 34

2 Multi-cellular Feature Representation Learning **35**

2.1 Introduction 37

2.2 Methods 39

2.2.1 Deep Variational Autoencoding Networks 39

2.2.2 MEMA Dataset . . . 40
2.3 Results . . . 41
2.3.1 VAE Analysis . . . 42
2.3.2 Latent Space Walking . . . 42
2.3.3 Measuring Organization with Human Annotation . . . 46
2.3.4 Characterizing Micro-environment Perturbation . . . 49
2.3.5 HCC1143 . . . 50
2.4 Discussion . . . 52
3 Neural Estimation of Metastatic Origin 55
3.1 Introduction . . . 57
3.2 Methods . . . 58
3.2.1 Data Set . . . 58
3.2.2 Learning Approach . . . 59
3.3 Results . . . 61
3.3.1 Quantitative Localization of Liver Cancer in Whole Slide Images . . . 61
3.3.2 Quantitative Whole Slide Image Classification of Metastatic Origin . . . 62
3.3.3 Clinical Benchmark Comparison Study . . . 65
3.4 Discussion . . . 67
4 DeepHAT: Deep Histology Annotation Tool 71
4.1 Introduction . . . 73
4.2 Data . . . 75
4.2.1 Whole Slide Images . . . 75
4.2.2 Whole Slide Manual Annotations . . . 75

4.3 DeepHAT Design 76
4.3.1 Whole Slide Image Preprocessing 76
4.3.2 Histological Representation Learning 77
4.3.3 Feature Manifold Projection 78
4.3.4 Interactive Annotation Application 79
4.3.5 DeepHAT Accessibility 81
4.4 Results 83
4.4.1 Single Slide Annotation 83
4.4.2 Tumor Annotation from Whole Slide Primary Tissue 84
4.5 Discussion 86
5 Primary-Metastatic Transfer Learning 91
5.1 Introduction 93
5.2 Methods 94
5.2.1 Computational Pipeline 94
5.2.2 Whole Slide Pre-Processing 95
5.2.3 Learning System Architecture 97
5.3 Results 97
5.3.1 Prior First-Stage Tumor Identification 97
5.3.2 Filtering Normal Primary Tissue 99
5.3.3 Primary-to-Primary Prediction 101
5.3.4 Primary-to-Metastatic Classification 105
5.3.5 Primary & Metastatic Feature Divergence 109
5.4 Discussion 112

6 Conclusion **115**

6.1 Limitations 115

6.2 Future Research Directions 118

6.3 Dei Ex Machinae 123

List of Figures

1.1 Biology across physical scale 3

1.2 The tumor microenvironment 4

1.3 Metastatic processes 7

1.4 Whole slide image 9

1.5 Detailed digital pathology perspective 10

1.6 The perceptron and MLP 13

1.7 Convolutional operation on 2D functions 16

1.8 Convolutional neural network 17

1.9 Backpropagation 19

1.10 Variational autoencoder 22

1.11 Convolutional operation and the human eye 25

1.12 Understanding convolutional filters 26

1.13 Deep learning work-flow for digital pathology 28

2.1 MEMA platform 41

2.2 VAE encoding of MEMA spots 43

2.3 t-SNE embedding of MEMA spots 44

2.4 Principal feature manifold 45

2.5 VAE feature density across MEMA annotations 47

2.6 PCA and LDA projections of MEMA encodings 48

2.7 Hierarchical clustering of micro-environment perturbations 51

2.8 HCC1143 MEMA projection 53

2.9 HCC1143 unsupervised clustering versus ligand subtype 54

3.1 Whole slide images of metastatic liver cancer 59

3.2 Deep learning approach for metastatic origin prediction 60

3.3 Inception v4 Architecture 61

3.4 NEMO tumor region identification results 63

3.5 NEMO metastatic origin classification results 64

3.6 Class-relevant example tiles 65

3.7 Example whole slide images across classification error types 67

4.1 QuPath annotation tool user interface 76

4.2 VAE projection of liver metastases 79

4.3 VAE tile projection of liver metastases 80

4.4 DeepHAT information and interactivity flow 81

4.5 DeepHAT user interface 83

4.6 DeepHAT on a single WSI 85

4.7 Histopathological feature space projection 86

4.8 DeepHAT vs. WSI feature manifold annotation 87

5.1 Transfer learning procedure overview 95

5.2 Tile and WSI sample counts 96

5.3 ResNet50 Network 98

5.4 Tumor filter comparison 99

5.5 Classification results for tumor identification in primary tumors 100

5.6 Tile and WSI classfication confusion matrices 102

5.7 Rank confidence of Metastatic Origin Classifier 103

5.8 Multi-class AUROC 104

5.9 Spatially-resolved predictions of metastatic origin 105

5.10 True class prediction distributions 106

5.11 Metastatic origin model AUROCs 108

5.12 Whole slide classification 109

5.13 tSNE-VAE embedding of primary and metastatic samples 111

5.14 KL divergence between primary and metastatic tiles 112

6.1 Automatic IHC panel recommendation engine 119

6.2 XAE architecture for unsupervised domain translation 124

List of Tables

2.1 ANOVA of VAE and hand-crafted features 49

3.1 Class-specific classification statistics 65

3.2 Misclassified slides by NEMO model and pathologists 66

List of Abbreviations

Abbreviation	Definition
ANN	Artificial Neural Network
ANOVA	Analysis of Variation
API	Application Programmer Interface
AUROC	Area Under the Receiver Operator Characteristics Curve
CA	Colonic Adenocarcinoma
CNN	Convolutional Neural Network
CPU	Central Processing Unit
DeepHAT	Deep Histopathological Annotation Tool
ECM	Extracellular Matrix
GAN	Generative Adversarial Network
GPU	Graphics Processing Unit
GIST	Gastrointestinal Stroma
HCF	Hand-crafted Feature
HDF	Hierarchical Data Format
H&E	Hematoxylin and Eosin
HMEC	Human Mammary Epithelial Cells
HPC	High-Performance Computing
IHC	Immunohistochemistry
KLD	Kullback-Leibler Divergence
LDA	Linear Discriminant Analysis
LINCS	Library of Network Cellular Signatures
MEMA	Microenvironment Micro-array
MEP	Microenvironment Perturbation
NC	Neuroendocrine Carcinoma
OHSU	Oregon Health & Science University
PFM	Principal Feature Manifold
PCA	Principal Component Analysis
ROC	Receiver Operator Characteristics
SGD	Stochastic Gradient Descent
SSIM	Structural Similarity Index
TME	Tumor Microenvironment
tSNE	t-Stochastic Neighbor Embedding
UMAP	Universal Manifold Approximation
VAE	Variational Autoencoder
WSI	Whole Slide Image

Acknowledgments

I am indebted to an great number of people for the opportunity to pursue this work over the past many years. For facilitating me towards and granting me the opportunity to join the computational biology program and biomedical engineering department here at OHSU, I would first like to thank Jackie Wirz, Adam Margolin, Paul Spellman, Monica Hinds, and Owen McCarty. In retrospect, joining this incredible group has been one of the greatest decisions of my life, and it would not have been possible without the commitment of these individuals to the importance of quality student mentorship.

I would next like to thank my dissertation advisory committee for their on-going support of both myself as a graduate student and this body of work. First, I'd like to thank Xubo Song for serving as the external member of this committee, and for providing my first comprehensive instruction in machine learning several years ago. I'd like to thank Andrew Adey in particular for his early support, encouragement, and warm console over the years; it's been a pleasure to stay in touch with the lab of my first research rotation as it has done so well over the years, even though its current graduate student crop is composed of rather rank amateur Magic players. I'd like to thank Guillaume Thibault for his instruction in computer vision, for his guidance in improving the quality of my work and the manner in

which I present it, and for his candid take on the people and events around our humble research community. I'd like to thank Chris Corless for his leadership of the immense complexity of the Knight Diagnostic Labs and for his enthusiasm to shake things up in an otherwise rather staid field; thanks for taking time off from boondocking to kick this thing off. I'd like to thank Laura Heiser for her leadership of my committee; for her piercing questions, insights, guidance throughout this work; and for her knack for making meaningful connections between ideas, technologies, and people. I am especially deeply thankful for and indebted to my primary mentor Young Hwan Chang for his support and guidance throughout the entire tenure of this work, for his gentle encouragement, and for his meticulous attention to detail; as his first graduating student, I have been humbled by the opportunity to learn along side you and witness the lab grow in number and in stature. Lastly, I am incredibly thankful to Joe Gray for making my environment possible. Your leadership in scientific team-building, sage wisdom in matters of life and science, and innovative thinking have impressed upon me the deepest during my time in graduate school; the chance to have had the opportunity to bear witness to the incredible strides being made in how cancer patients are treated has been a singular privilege of a lifetime. I am lucky to have had an immensely satisfying PhD experience during this incredibly exciting time in biomedical imaging research, and I owe the satisfaction of it all to these, my closest mentors. Thank you.

I'd like to thank the many other individuals who contributed to my early research rotations and who provided valuable perspective on life as an academic researcher inside and outside the lab, specifically I'd like to thank Uchenna Emechebe and Lincoln Shenje for early lessons in academic life; Lucia Carbone and Shawn Chavez for early exposure to

genomics research; Devo Goldman and Harv Flemming for the opportunity to rotate in the Oregon Stem Cell Center; and Guanming Wu for his much-needed support when it mattered most. I'd also like to thank members of my qualifying exam committee Jeremy Goecks, Jim Korkola, and Fay Horak for their support in crystallizing research ideas. For keeping myself on track and on time, I must thank the many helping hands of incredible administrators, especially JoAnn Takabayashi, Holly Chung, Lauren Kronebusch, Kattie Crossen, and Erica Regalo for their support along the way. I have been fortunate to have had incredible resources at my disposal during this work, and so I am indebted to the on-going support from Jonathan Jubera, Anne Carlson, Paul Heinlein, Marion Hakanson, Nathan Smith, and Aletha Lesch for their various roles in supporting intellectual property development, advanced computing infrastructure, and the Knight BioLibrary.

I'd like to thank the many friends and scientific collaborators for making my time here educational, exciting, and fun. Especially I'd like to thank Mark Dane, Elmar Bucher, Jim Korkola, Zuzana Taratova, Reuben Hoffman, Nick Calistri, Jenny Eng, Koei Chin, Jessica Reisterer, Kevin Loftis, Kevin Stoltz, Jeremy Copperman, Zeynep Sayar, Ece Eksi, Ellen Langer, Tub Anekpuritanang, Ying Wang, and Hassan Ghani. My life as a graduate student would not be the same were it not for my good friends and close confidantes that I've had the chance to meet along the way, especially Mike Soroka, Kristof Törkenscy, Eileen Torres, Casey Thornton, Brit Alperin, Ryan Mulqueen, Ben Cordier, Julianne David, and Kristen Stevens. I also wish to thank my childhood friends Nat Steinsultz and Joe Kopecky who preceded me through the PhD gauntlet, and for giving me confidence that if they could do it, anyone can. Also I'd like to thank my dear friends Mykal Mantyla and Sam Mortensen who, through giving me the opportunity to share some of the work being done

by this research community during their own time of great uncertainty, have given this work greater meaning to myself. I especially wish to thank my fellow lab mates Elliot Gray, Luke Ternes, Tina Ghodsi, and especially Erik Burlingame for all of their feedback and meaningful contributions over the years. Cheers to the many productive discussions of science, and many beers, had in varying states of sobriety. It's been a fun ride with you all.

Outside of the lab, my life has been enriched with deep goodness over the past several years. For centering me and providing the essential comforts of home during times of both successes and failures, I want to thank Marvin the cat for keeping my lap warm during many days and nights of coding and writing, and Iris for her enduring love, support, and patience throughout everything. During these unprecedented times of human isolation, I am beyond grateful and fortunate to have had so many homely comforts that have kept me satisfied and fulfilled. Of course, I wish to thank both my parents for their many gifts of education, opportunity, life, and love over the years. Both first-generation college graduates themselves, their lessons in the importance of good education and wise investment decision-making have contributed greatly to my choice to continue my education and pursue a PhD; to them I dedicate this work.

Finally, I am profoundly indebted to the many patients, both current and passed, whose biology, having been made available through various research repositories around the world, has contributed to this research. May this work do their many sacrifices justice, and may their gifts to science be not forgotten.

Abstract

Controlling cancer requires comprehensive understanding of the molecular, cellular, and organizational properties of tumor tissue. While clinical pathology has served as a gold standard for cancer diagnosis for over a century, the field continues to largely rely on visual inspection of sectioned and stained tissue under the microscope by expert pathologists. Emerging technological advancements in scanning equipment have contributed to wide-scale digitization of whole slide images with clear clinical connotation, while parallel strides in artificial intelligence have enabled computational models of computer vision shown to meet or exceed human capabilities for identifying features of interest and abnormalities in imaging data. This work integrates deep learning systems for histopathological image analysis to quantitatively and qualitatively evaluate spatial characteristics of tumor biology to better guide clinical diagnosis and treatment of cancer. First, I illustrate an unsupervised approach to learn meaningful representations of multicellular organization both in response to environmental perturbation and across diverse histopathological tissue types. Second, I demonstrate a supervised approach that enables clinical-grade inference of the metastatic origin of secondary cancer from whole slide histological sections of liver metastases. Third, I introduce a semi-supervised annotation tool that retains an expert human pathologist in the

loop to resolve a critical research bottleneck in digital pathology by accelerating whole slide annotation. Finally, I present a learning approach that transfers learned spatial features from images of primary cancers to infer the origin of secondary metastatic cancers. These studies support the application of emerging systems for computer vision, artificial intelligence, and interactive manipulation of biological data to accelerate pathological evaluation processes and augment human understanding of biology and disease.

Chapter 1

Introduction

I believe the treatment is an absolute cure for all forms of cancer.

Dr. J.E. Gilman,
on X-rays (1901)

But what... is it good for?

Computing Systems Engineer, IBM
on the microchip (1968)

A convergence of recent advancements in medical imaging instrumentation, clinical oncology, and artificial intelligence have given rise to the rapidly evolving field of digital pathology, which broadly investigates the application of advanced computational models of computer vision to biomedical imaging data. This chapter introduces key concepts and recent findings of spatial biology of cancer, clinical pathology, and artificial intelligence that give context to the remainder of this work.

Section 1.1 introduces the field of spatial systems biology and key technologies that have enabled the deep profiling of spatial characteristics that have contributed to enhanced understanding of tumor function. Section 1.2 introduces key concepts underlying the field of computer vision and deep learning, which have undergone rapid developments in recent years. Section 1.3 provides a research landscape overview of recent approaches to

incorporate artificial intelligence learning systems to biomedical image analysis. Section 1.4 concludes with a summary of current opportunities and challenges in the field of digital pathology, and section 1.5 summarizes the contributions of this work and provides an overview of the remaining subsequent chapters.

1.1 Spatial Systems Biology of Cancer

Our understanding of biology canonically fits into a hierarchical teleology that is typically organized along an axis of physical scale. Figure 1 illustrates the scale of living systems from atoms to molecules up through proteins, protein complexes, structures, cells, tissues, organs, individuals, and out into multi-agent ecologies. While biological research has historically been a largely reductionist enterprise - breaking down complex systems to study constituent units - an emerging holistic view seeks to synthesize biological findings into a more comprehensive context through the process now known as systems biology. Cancer systems biology extends this approach to tumor biology to better understand how interactions between signaling mechanisms, cells, and microenvironmental factors contribute to tumor growth and resistance to therapy. By elucidating mechanisms by which tumors develop resistance to targeted therapy, treatment strategies may be modulated to ensure durable reponses both to improve and extend quality of life.[1]

Findings stemming from this approach have compelled some researchers to consider cancer as a systems biology disease, characterized by the disregulation of interacting elements intended to maintain normal physiological homeostasis.[3] Cancer itself is not one disease, but an umbrella covering physiological conditions that exhibit characteristic hallmarks such

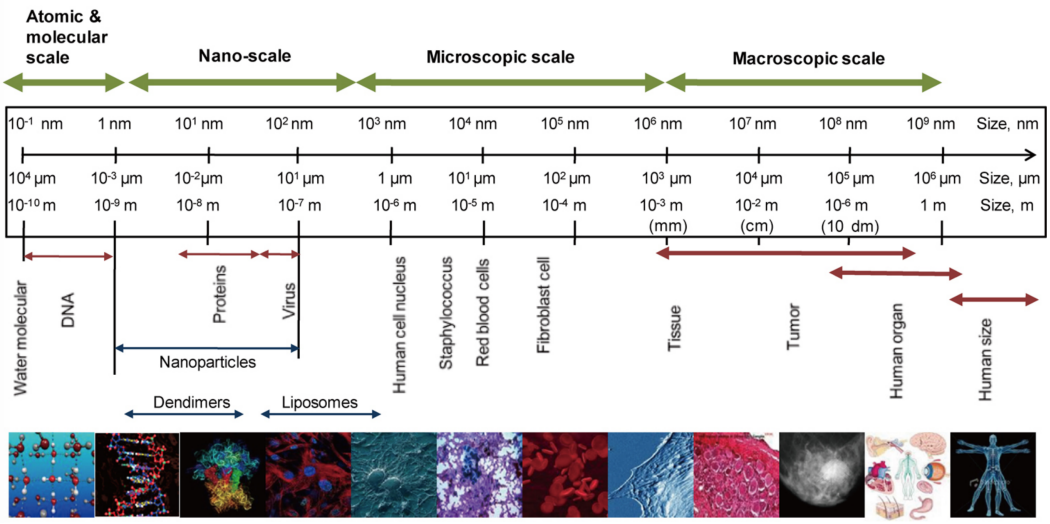


Figure 1.1: Biology across physical scale. The role of the pathologist generally focuses on scales ranging from cells through tissue and tumor. Figure reproduced from Kim, 2013[2]

as active invasion and metastasis, sustained proliferative signaling, growth suppression invasion, and cell death resistance.[4] As tumor tissue undergoes uncontrolled cellular proliferation, tissue organization necessary to sustain normal homeostasis is disrupted. The emergent complex architecture composed of specialized cellular subtypes is generally referred to as the tumor microenvironment.[5] Figure 1.2, reproduced from Hanahan and Weinburg,[5] illustrates the multi-cellular complexity in spatial systems architecture presented by tumor cells, non-tumor cells, immune cells, endothelial cells, fibroblasts, and a host of other specialized units. Emerging appreciation for how cells interact in complex environments poses a new challenge to understand how spatial abberations and the microenvironment collaborate to disregulate cell differentiation, proliferation, apoptosis, and motility.[6]

1.1.1 Mechanisms by Which Cancer Arises

The mechanisms by which cancer arises are diverse, complex, and in many cases poorly understood. Cancer is believed to typically arise from a single cell that acquires one or

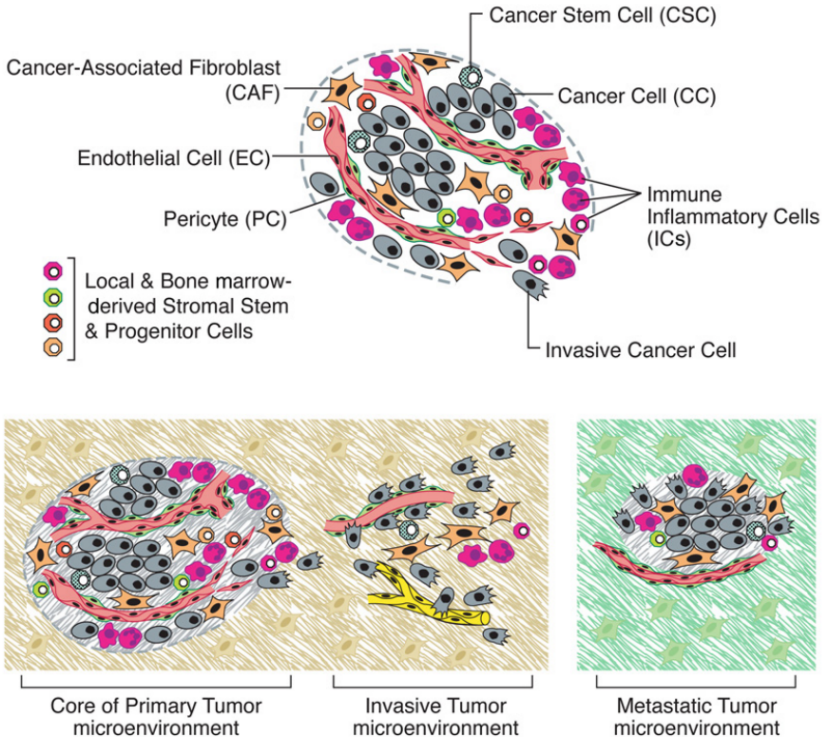


Figure 1.2: The tumor microenvironment is composed of diverse and interacting constituent components. Reproduced from Hanahan, 2011[5]

more mutations that impacts its ability to regulate self-replication, and ultimately induces cellular immortality, which results in unconstrained proliferative growth. As a mutated cell divides, it passes on its defective properties to its progeny, accelerating the outgrowth of the now cancerous cells. The process by which cancer arises is broadly defined as tumorgenesis. A single cell may acquire immortality through a unique set of steps, but causal mechanisms tend to point towards a spontaneously acquired mutation in one or more genes associated with cancer promotion or in one or more genes associated with cellular self-termination. For example, the RAS family of proteins encoded by the RAS genes are involved in intracellular signal transduction which promote cell growth when activated, such that a mutation causing RAS to become permanently active may induce cancer by forcing cells to continuously

divide.[7] Alternatively, the protein encoded by the p53 gene is known to regulate apoptosis such that a mutation that interrupts normal function of p53 may remove a primary mechanism by which cells normally halt their own replication.[8]

Tumors likely develop over several years as cells begin to divide at a higher rate and acquire additional genetic mutations and epigenetic alterations over time. Even though a cell that has been immortalized might begin the process of unconstrained division, other physiological stresses work against cancer's growth by imposing limitations on available resources, such as oxygen; while cancers are known to upregulate oxygen supply through promotion of angiogenesis, a cell that runs out of oxygen will eventually die, whether the cell is cancerous or not.[9] The telomere structures that cap the terminal ends of chromosomes similarly provide an intrinsic limitation on cell divisions, as a cell whose telomeres become depleted can become non-viable after additional replications, though upregulation of telomere-repairing telomerase is known to circumvent this control.[10]

Because most cells contain two copies of each chromosome, the two-hit hypothesis in the context of tumorgenesis postulates that mutations must be aquired in both promotion of oncogenes and of inactivation of tumor suppressor genes in both alleles in a cell.[11] Whether cells may be spurned to induce replication by over-expression of oncogenes or whether the intrinsic rate-limiting mechanisms of tumor suppressor genes malfunction, only a single cell is required to initiate tumorgenesis and, eventually and over many generations, result in a cancer diagnosis.

1.1.2 Metastasis

Although primary tumors themselves present a significant biological burden, metastasis remains the cause of 90% of deaths from solid tumors.[12] In an overwhelming majority of cancer patients, a diagnosis of metastatic disease is strongly associated with terminal illness.[13] Because the complex microenvironment of a tumor affects its ability and propensity to metastasize to other sites in the body,[14] targeting metastatic signalling processes may present unique opportunities to improve patient outcome.[13]

The process of metastasis is complex, involving a series of physiological and physical steps necessary for a cancer cell to extravasate outside of the primary tumor, survive circulation, seed a target organ, and engage in persistent growth.[15] This process is shown in Figure 1.3 which illustrates the physical forces necessary to disrupt normal physiological function in such a way as to transport cancer cells with metastatic potential away from the primary tumor and into a secondary site with suitable microenvironment conditions that enable proliferative growth. Although metastases can arise from virtually any primary tumor in the body, specific locations exhibit preferential hosts for metastasis, particualrly brain, lung, bone, and liver.[16]

1.1.3 Imaging Science

Characterizing complex tumor architecture requires methods capable of profiling molecular physiology of tissue while preserving the spatial context of those measurements. Considering perhaps that humans generally perceive their environment through visual analysis of their surroundings, for centuries imaging has provided a compelling medium to understand tumor tissue, and remains a foundational approach to unravel the spatial structure of biological

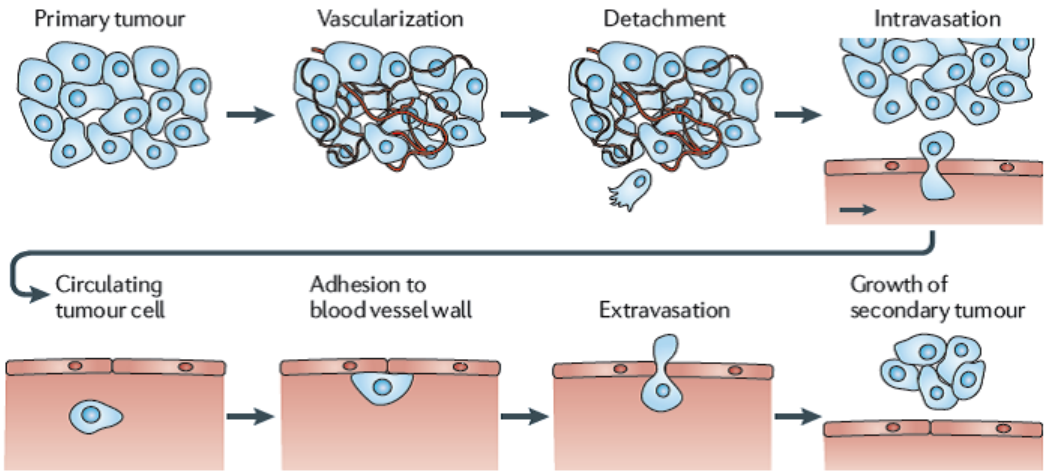


Figure 1.3: Metastases are distal outgrowth colonies of cells dispatched from a primary tumor through blood or lymph vessels or by extravasation through solid tissue.

systems.[17] Unlike most genomic analyses which destroy physical compartmentalization during preparation, imaging science enables visualization of spatial interactions of a wide variety of physical components in the tumor microenvironment at single-cell resolution.[17] An array of imaging technologies have elucidated biological mechanisms operating at vastly different scales, from nanoscale measurements made using quantum dot imaging all the way through meter scale measurements made with PET (Positron Emission Tomography) and MRI (Magnetic Resonance Imaging) technologies.[18]

Clinical diagnosis of cancer typically requires a surgical resection or biopsy of the secondary solid tumor. Once resected, the sample is chemically fixed and embedded into a solid medium for storage and handling. From the resulting block, thin sections of tissue are sliced, placed on a slide, stained, and viewed under a microscope. Despite generally consistent methods for histopathological specimen preparation, clinical pathologists must be flexible to technical artifacts and general uncertainties in discriminating between different cancer types. Technical challenges with tissue preparation also manifest in digitized scanned

images, including tissue tearing, tissue folding, and inadvertent capture of unwanted debris, while staining artifacts due to inconsistent preparation or application of the H&E stain are also not uncommon, often leaving regions of tissue either under- or over-stained. Variability both between pathologists as well as within a single pathologist across different settings are referred to as inter-operator and intra-operator variability, respectively, and remains a significant challenge both in research and clinic.[19–22]

Emerging multiplexed imaging technologies afford an entirely different approach for tissue analyisis. Multiplexed immunohistochemistry,[23,24] cyclic immunofluorescence,[25] multiplexed imaging mass cytometry,[26] and CODEX[27] have all emerged in the past several years, each employing different fundamental mechanisms to elucidate the presence or absence of molecular biomarkers of interest while retaining their spatial organization in whole tissue. Despite an emerging generation of exciting biomedical imaging technologies, a vast majority of clinical pathology still relies on hematoxyin and eosin (H&E) staining. The protocols for generating H&E images are well-established, providing a cheap and easily accessible method for producing high-contrast representations of tissue capable of displaying a broad range of cytoplasmic, nuclear, and extracellular features.[28] An example whole slide image (WSI) is shown in Figure 1.4. At low magnification, the tissue appears similar to what might be seen with a human eye. However, while microscopes are essential for discerning fine details of a tissue composition, digitization of such images captures detailed structure in a single image file that may be stored, transported, and viewed in a computer.

Increasing magnification of a whole slide image illuminates far greater detail than what can be seen at whole-slide level. Figure 1.5 illustrates a zoomed-in dramatic example of an H&E stained section of an invasive adenocarcinoma (left half of the image) invading normal

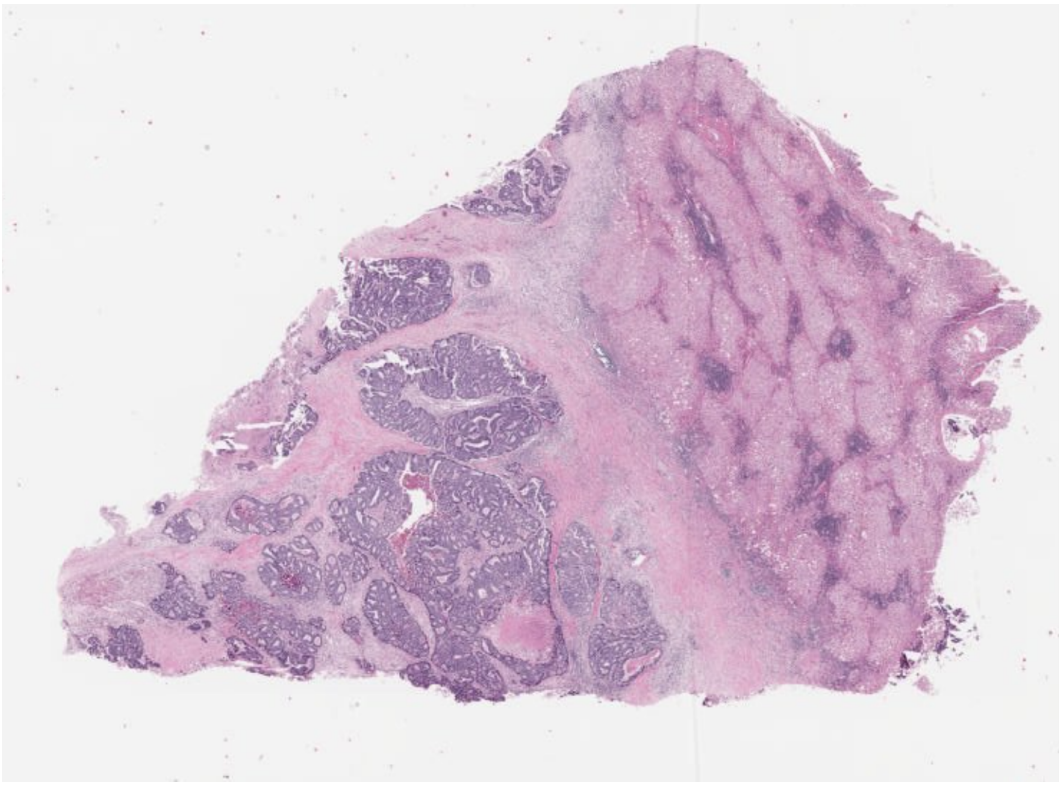


Figure 1.4: A digital whole slide image of tissue sectioned and stained with hematoxylin and Eosin reveals remarkably diverse spatial structure within living systems. Seen from a low-magnification view, distinct regions of the sample are evident. The blue color corresponds to the hematoxylin staining cellular nuclei while the pink color corresponds to the eosin staining the extracellular matrix and cytoplasm. Together, H&E staining provides a reliable and high-contrast representation of tissue features that are essential for guiding a pathologist to an accurate diagnosis.

liver tissue (right half of the image), each separated by a barrier of immune infiltrate. Careful inspection of these images by a trained pathologist are generally sufficient to establish preliminary diagnoses, and even today are generally considered a gold standard of cancer diagnostics.

Although whole slide histology has historically relied on visualizing stained tissue under a microscope,[29] instrumentation capable of scanning whole slide images at high resolution has enabling capture and digital storage of histopathological specimens.[30] Since their first

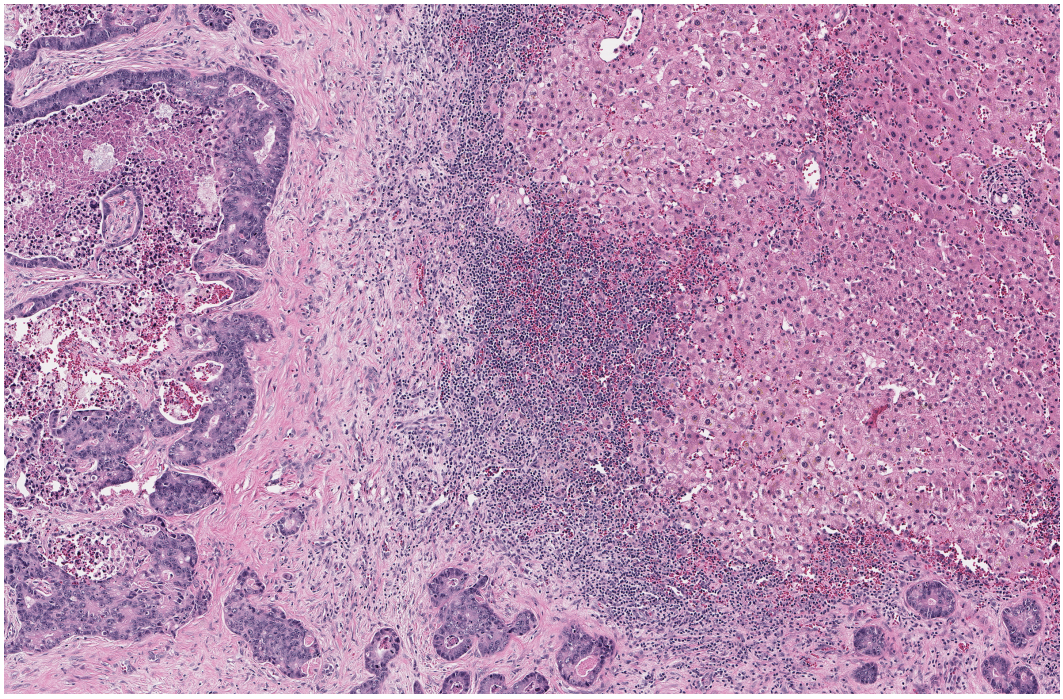


Figure 1.5: Digital pathology enables computer-based interaction with high-resolution images of scanned tissue sections, such as this example of invasive metastatic adenocarcinoma of the liver. At this resolution and magnification, individual cellular nuclei become clearly evident while the texture and shape of the cellular cytoplasm enables a pathologist to readily classify cells according to their type. For example, the broad, regular shapes of the large cells at the right of this image are easily classified as normal liver tissue, while the small dense cells with little cytoplasm near the middle of this image are characteristics of immune cells. The disorganization of tissue seen at the left side of this image is a characteristic property of many cancers.

introduction in 1990, whole slide scanners have undergone significant technological progress, and have enabled digitization of whole slide images accessible to many clinics.[31] While the field of pathology has at times been slow to adopt technological advacements, efforts to incorporate digital whole slide imaging have gained traction in pathology departments for educational, diagnostic, and research purposes, in some instances even replacing traditional glass-slide microscopy in clinical practice.[32]

In summary, the current state of oncology is confronted with three features of interest relevant to this work: the abundance of imaging data capturing the context of a tumor's

microenvironment, the pressing clinical need to establish durable response to therapy, and persistent concerns regarding inter-operator variability of tissue inspection. Computer vision systems have, in recent years, demonstrated success in establishing high-level understanding of tissue, as will be discussed in Section 1.3, but to better understand how and why computer vision systems have undergone extraordinary progress in the past decade requires a brief introduction to the core computational features of modern computer vision systems.

1.2 Learning Systems Architecture

As pathology migrates into a digital medium, it brings with it tremendous opportunity. The abundance of consistent data coupled with health care records, co-morbidities, health outcomes, and response to therapy presents a rich space to test and validate biological hypotheses and design new tools. Learning to associate features of histopathological tissues specimens with targets or labels of clinical interest requires a computational approach capable of learning spatial features contained within whole slide images that are relevant for a clinical task; the broad success of deep learning and artificial neural networks has opened a promising avenue to begin addressing key problems in digital pathology.

1.2.1 Artificial Neural Networks

Although the terms neural network and artificial intelligence have recently come back into vogue, the origins of neural computing architecture extend as far back as the 1940s.[33] with the introduction of the McCulloch-Pitts model of the neuron. This concept was the first effort to formulate a mathematical abstraction of neuronal function by explicitly defining

system input, transformation, and output of a self-contained system modeled after biological information processing.[34] The extension of the mathematical abstraction of the McCulloch-Pitts model to a typical classification task was first demonstrated by Frank Rosenblatt in 1960 through a design he called the perceptron.[35] Though a relatively simple algorithmic device, the perceptron, as shown in Figure 1.6, remains the foundational basic building block of neural network systems. In its most basic form, a perceptron is composed of a set of weights $[w_1, w_2, ..., w_n]$ corresponding to each input feature $[x_1, x_2, ..., x_n]$ plus a bias weight w_0. The perceptron modeled as function $n(x)$ computes the dot product between inputs and weights, adds the bias term, and passes the result through some activation function σ, which presumes a non-linear function such as the signum, or sign function, though many alternatives exist, giving the fundamental perceptron function

$$n(x) = \sigma\Big(w_0 + \sum_{i=1}^{n} x_i w_i\Big) \tag{1.1}$$

Modern neural network architectures combine thousands or even millions of similar types of computing units together to give rise to extraordinary computational complexity.[36] Because the fundamental perceptron unit takes as input a set of multiple features and returns a single output, perceptrons can be stacked both vertically and horizontally to create networks with greater capacity and depth. In modern terminology, a set of perceptrons each processing the same input data but with unique sets of learned weights is referred to as a *layer* in which the number of perceptron nodes is generally referred to as the *width* of a layer. A model that stacks multiple layers in series is generally referred to as a multi-layer perceptron, in which the number of composed layers is generally referred to as the model's

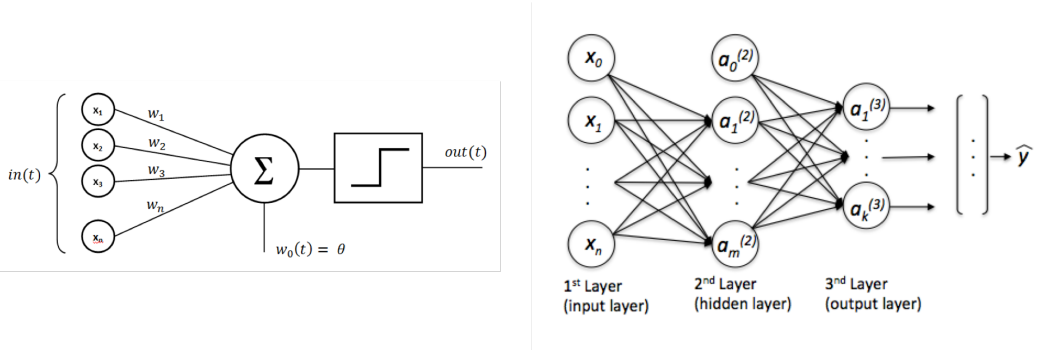


Figure 1.6: (Left) The perceptron is the fundamental building block of modern artificial intelligent systems by multiplying input data against a set of learned weights and passing the resultant sum through a non-linear activation function. (Right) Stacking perceptrons both in height and width produces a basic multi-layer perceptron, the most basic neural network architecture.

depth, as shown in Figure 1.6. Computational resources afforded with modern graphical processing unit hardware enable training of neural networks of significant width and depth whose complexity and performance scales with the overall size of the network, giving rise to the term *deep learning* to refer to such models that contain more than one hidden layer in the network.

1.2.2 Convolutional Neural Networks

Computer vision broadly refers to the computerized processing of 2- or 3-dimensional imaging data for the purposes of extracting or inferring a high-level understanding content within an image.[37] Although humans generally perceive the world in 2 and 3 dimensions, an artificial neural network model assumes independence amongst its input features, treating them independently and summing up their marginal contributions. Human perception relies on the spatial relationship between individual measurements in an image, such that neighboring pixels provide relative context to their neighbors and vice versa. The use of

convolutional neural networks began with seminal work by Yann LeCun in 1988 in which a small convolutional neural network was optimized to recognize hand-written digits[37] and was reignited in 2012 when a team led by Geoffrey Hinton surpasses state-of-the-art results on the ImageNet dataset classification challenge using a convolutional neural network that had been optimized on graphical processing unit processing hardware.[38] Since then, the use of CNNs has dominated recent imaging processing literature, having accelerated advancements across the full spectrum of image and video processing applications, motivating recent advancements in object detection, scene transformation, and semantic segmentation. Understanding how and why these types of neural architecture routinely achieve state-of-the-art results requires a brief overview of the mathematics behind their operation.

In analog signal processing, the convolution operation refers simply to flipping and sliding two functions $f(t)$ and $g(t)$ against each other by some step τ and computing the integral of the resultant function.

$$s(t) = \int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau \tag{1.2}$$

Intuitively, convolution tells us how a system might respond given an input stimuli $f(t)$ and a known response of the system to some unit step or impulse function $g(t)$ as the stimuli passes through the system. The *flipping* of a signal ensures that the earlier time-step of the signal is passed through the system first. Under the assumption of a linear time-invariant system, additive stimuli are compounded within the system response as the two functions are convolved against each other.

In the context of deep learning, the convolutional operator is considered a multi-dimensional

and discrete-space function in which an input image I of dimensions M is convolved against a learned kernel K of a chosen size to produce an output feature map S.‘ A visualization of a simplified case is provided by Ian Goodfellow and reproduced below in Figure 1.7.[39] To extend the illustrative example, consider an input and kernel $K \in \mathcal{R}^{m \times n}$, each with two dimensions in which the corresponding feature map is defined by

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n) \tag{1.3}$$

Because the convolution operation is extensible to any number of dimensions, a stack of kernel layers can be treated similarly to a set of nodes in a layer of a neural network. The fundamental convolutional neural network is generally composed of a few layers of convolution kernels followed by one or more densely-connected layers in which the number of output nodes generally corresponds to the number of unique classes that a classification task might occupy. Figure 1.8 illustrates the general principle, in which an input image is convolved with the learned kernel weight matrix to generate stacks of feature maps corresponding to the chosen depth of each kernel. Depending on the task, the feature maps may be flattened and passed through a set of fully-connected layers to generate predictions, such as an image's class given the learned feature map representations. However, other learning tasks such as segmentation retain multi-dimensional feature representations. Also shown in Figure 1.8 are *pooling* layers, which are designed to collapse the output feature maps typically by some summary statistic, such as max pooling,[40] to help regularize the learned representations and make the model more invariant to small changes in input. By defining an appropriate metric of success for the vision task, such as accuracy for classi-

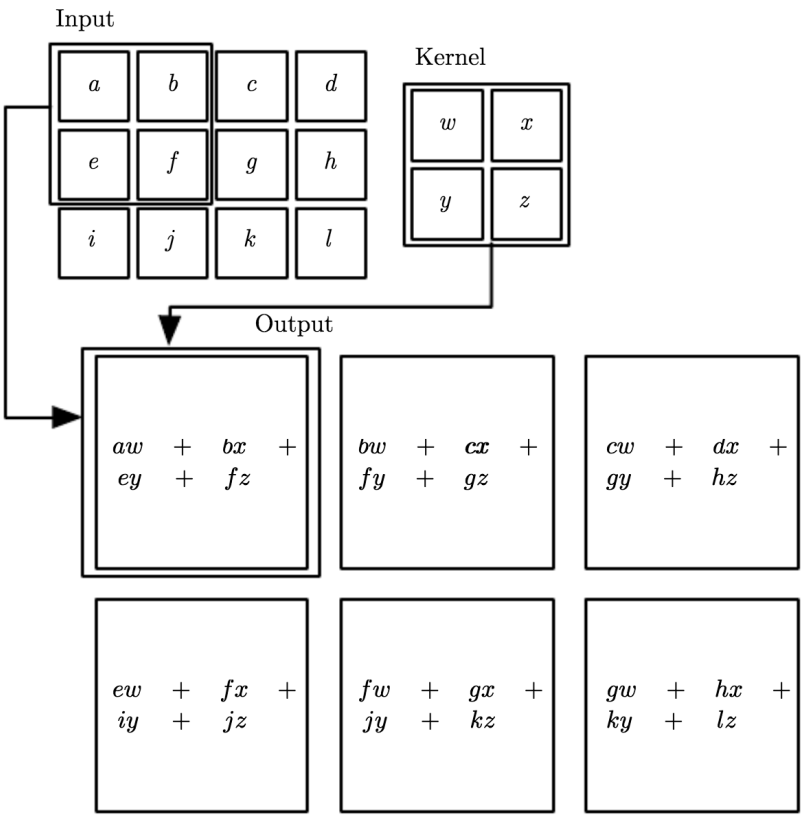


Figure 1.7: The convolutional operation as applied to 2D discrete functions. For each discrete step size, the running product sum of overlapping elements in the input and kernel are computed as a new 2D output map. This example illustrates an input composed of a single-channel, but the method is readily extensible to inputs of arbitrary depth. Figure reproduced from Goodfellow, 2016[39]

fication or intersection-over-union for segmentation, error is back-propagated through the convolutional kernels to update their weights to improve feature learning relevant to the learning task.

The purpose of this type of learning system is to learn spatial features with the convolutional layers to generally optimally infer some class label y given some data point x. When learning a multi-class system to correctly predict a single y from x, it is often desirable to enforce a unitary summation constraint using the softmax activation function such that the

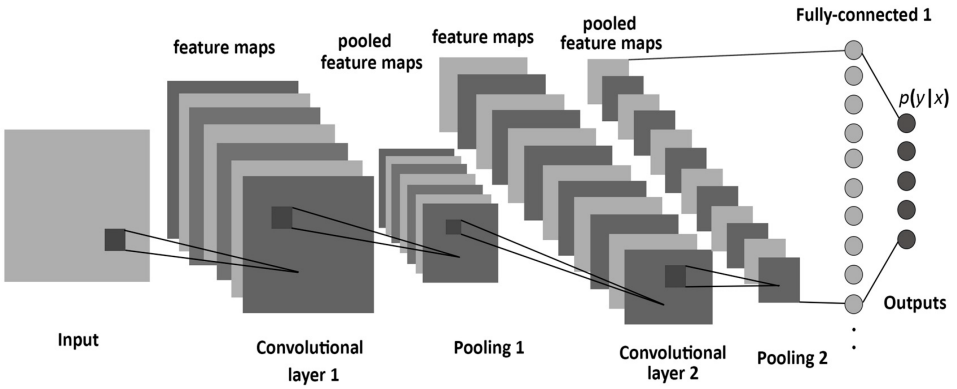


Figure 1.8: The convolutional neural network applies a sliding window across an input image. Because the weights of the window are held constant as it moves throughout the image, these networks are both efficient and robust. By learning two-dimensional functions, these types of layers extract spatial patterns within images to make predictions.

sum of all values from the layer add up to one, which enables a valid probabilistic interpretation of the resultant predictions. The standard softmax equation is shown in Equation 1.4 in which $\sigma : \mathbb{R}^K \to \mathbb{R}^K$ is defined for $i = 1, ..., K$ and $\boldsymbol{z} = (z_1, ..., z_K) \in \mathbb{R}^K$.

$$\sigma(\boldsymbol{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{x_j}} \tag{1.4}$$

The loss function between the training data labels represented by one-hot vectors and the distribution of class predictions provides an intuitive measurement of error by the classifier. Although classification is a foundation method in supervised machine learning, the use of convolutional layers across images can provide informative unsupervised representations as well. The variational autoencoder (VAE) learning architecture is employed in various subsequent chapters, and so a brief description of its design and function are provided.

1.2.3 Updating Learning Models

In either case of supervised or unsupervised learning with convolutional neural networks, the error measured by any metric must be utilized to update the model in such a way as to minimize the error during training. In the case of learning a classification model, the model's parameters θ compute a distribution of the probability $p(\boldsymbol{y}|\boldsymbol{x};\theta)$ of some target value $\boldsymbol{y}$ given some input $\boldsymbol{x}$. Optimizing the model parameters $\boldsymbol{\theta}$ requires first estimating an error measurement between the model's predictions and their ground truth targets, and then utilizing that error to update the parameters in such a way as to lower the estimated error during training. In the context of neural networks, passing a piece of data into a network and computing an output is known as a feed-forward operation, while using the error to update the weights of the network is known as backpropagation and is a key feature in how neural networks are trained efficiently.

Utilization of the back-propagation method[41] was first applied to artificial neural networks and image analysis by Yann LeCun in the late 1980s[42,43] and continues to underly the optimization of most neural network architectures. The backpropagation algorithm leverages the familiar chain rule of calculus and is best visualized as a computational graph such as the example shown below in Figure 1.9. This simple illustration shows that backpropagation of error can be achieved in a step-wise manner going backwards through a network starting at the output and evaluating how each of the weights of the model contributes to the total error of the loss function. By stepping backwards and nudging each weight of a layer either up or down in whichever direction that improves the model's performance, the model learns to reduce its loss function with each iteration. Thanks to the chain rule,

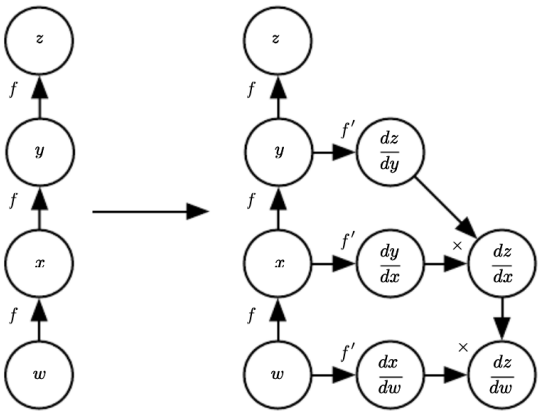


Figure 1.9: The backpropagation method for updating a neural network shown as a computational graph model in which each node represents a layer in an artificial neural network. In this example, we wish to understand how output z relates to operations at other steps in the computational graph $\frac{\partial z}{\partial w}$. The chain rule of calculus illustrates how the partial derivative may be factored out by computing the partial derivatives of each layer with respect to connected nodes. Adapted from Goodfellow, 2016.

this is done efficiently between each layer rather than needing to explicitly compute the contributions of each weight to the final output.

The backpropagation algorithm underpins the mechanism by which weights are typically updated in a network given some supervisory signal by descending along the error gradient with respect to the parameters of the network.[44] This process, known as stochastic gradient descent (SGD), refers to breaking a dataset into batches and updating the model over each batch instead of the entire dataset. The objective of stochastic gradient descent is to update the model parameters w over multiple batches of the dataset at each time step i by computing the gradient of the error function $Q(w)$ and updating the model weights by some learning rate or step size η.

$$w_i = w_{i-1} - \eta \nabla Q(w_{i-1}) \tag{1.5}$$

By iteratively backpropagating error to update model parameters with respect to error over multiple iterations through some optimization policy such as stochastic gradient descent[45] or adaptive learning policies such as the adaptive moments (Adam) optimizer,[46] neural network architectures arrive at some terminal parameterization of w that ideally achieves a global minimization of the chosen loss function. While backpropagation and SGD remain the *de facto* standard for modern neural network optimization, methods to achieve improved outcomes and avoid local minima continue to evolve.[47]

The Variational Autoencoder

While inferring the posterior probability $p(y|X)$ that a data point X belong to a given label y is generally considered to be a classification type problem, often the absence of training labels y necessitates unsupervised learning within the dataset under consideration. Many typical feature extraction methods are sensitive to subtle shifts in pixel offset when each pixel is treated independently, particularly when high-frequency content is present. This is particularly true in imaging data, where a feature extraction method should be invariant to the position, scale, or rotation of an object in a scene. In this case, rather than learning axes of variation within pixel-wise correlations across a dataset, latent features defining the content of an image are desired.

Autoencoders are a family of learning models that seek to learn optimal compression of data subject to an information bottleneck constraint by training both an encoder and a decoder simultaneously and minimizing the difference between an input data point and its decoded, encoded reconstruction. One significant challenge of basic autoencoders is their propensity to memorize data, due in part to the broad numeric space models can lever-

age to embed data representations.[48] The variational autoencoder (VAE) seeks to address challenges pertaining to memorizability by learning latent features through variational inference.[49] Bayesian inference methods seek to learn latent or hidden representations of data given observations made across a dataset by inferring the latent variables z given observations made in the data. By defining an encoder network parameterized by ϕ and a decoder network parameterized by θ, the learning objective of an autoencoder, shown in Equation 2.1, is to maximize the probability that a data point x_i is generated given a latent representation z while minimizing the divergence between the distribution of the latent variables with respect to an assumed prior $p(z)$

$$\mathcal{L}_i(x_i, \theta, \phi) = -\mathbb{E}_{z\sim q_\theta(z|x_i)}[\log p_\phi(x_i|z)] + D_{\mathrm{KL}}(q_\theta(z|x_i)\|p(z)) \tag{1.6}$$

Where D_{KL} is the Kullback-Leibler Divergence, defined for two probability distributions $P(x)$ and $Q(x)$ across propbability space $\mathcal{X}$ as and in which the distribution of the latent variable is typically defined as the normal Gaussian distribution $p(z) \sim \mathcal{N}(0, 1)$.[50]

$$D_{\mathrm{KL}}(P\|Q) = \sum_{x\in\mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right) \tag{1.7}$$

The general architecture of a variational autoencoder is shown in Figure 1.10 in which the model is composed of probabilistic encoder and decoder functions that compress and decompress data, respectively, through latent vector $\boldsymbol{z}$.

In imaging, we model both encoder ϕ and decoder θ as convolutional neural networks that can learn spatial features of interest in learning optimal latent representative features. Because the error of the variational autoencoder learning architecture relates to the *dis-*

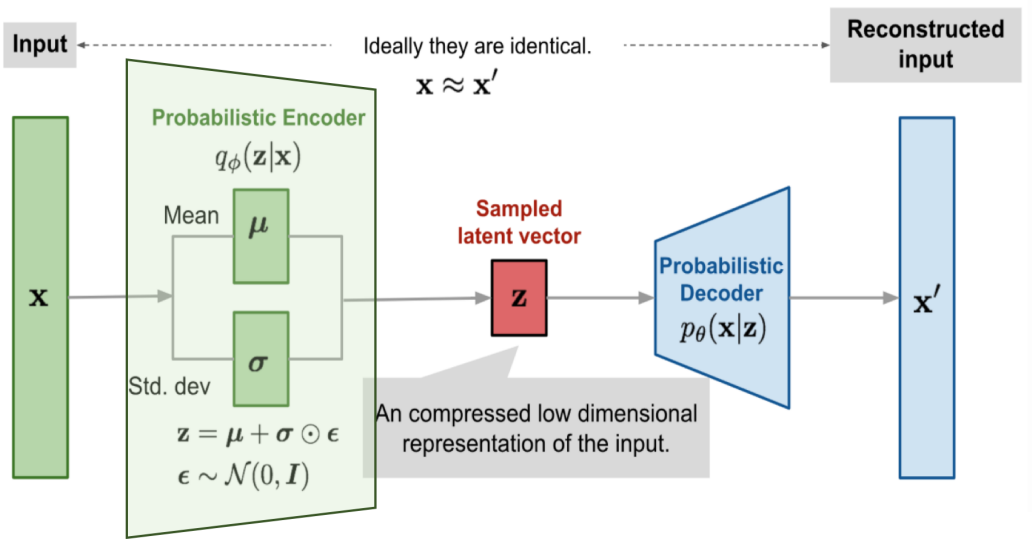


Figure 1.10: The basic design of a variational autoencoder learns latent representations of input data as $\boldsymbol{z}$. Modified from `lilianweng.github.io`

tribution of latent features, special consideration is required to ensure that changes to the model during training corrects for this effect. Although in principle, the distribution of $\boldsymbol{z}$ could be computed by sampling the prior expected distribution of latent features, the sampling operation is non-differentiable and can not be used to explicitly update the weights of the network. The *reparameterization* trick diverts the non-differentiable portion of the operation out of the network by parameterizing the mean and standard deviation of the latent distribution $\boldsymbol{z}$ and scaling the variance by a random variable ϵ. This simple transformation enables a fully differentiable network to infer the latent variables of a dataset subject to the information carrying capacity constraints of $\boldsymbol{z}$, given a dataset.

The VAE loss term is generally composed of two elements: the first penalizes reconstruction error and the second penalizes feature distribution. While the KL divergence is generally suitable for measuring statistical dissimilarity between two distributions, the other quantitative metric of interest for the purpose of unsupervised feature learning with a

variational autoencoder is the choice of dissimilarity metric between an input image and its associated reconstruction. Common metrics include the cross-entropy and mean squared error between an input and its reconstruction. The structural similarity index (SSIM), shown in Equation 1.8, is a simple formulation that has been shown to closely associate with human perception, such that high SSIM corresponds to high likelihood that a human will be unable to distinguish one image from the other.[51] By applying fixed window sizes across regions of an image, the SSIM measures evaluates the contributions of the mean intensities μ_x and μ_y, variances σ_x and σ_y, and covariance σ_{xy} of two images x and y and where c_1 and c_2 are small scalers added for numeric stability. While the choice of reconstruction error depends on the nature of the problem, the relative contribution of each term to the total error function is typically controlled by a balancing hyperparameter such as that used in the βVAE model.[52]

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2\sigma_y^2 + c_2)} \tag{1.8}$$

Although VAEs have been widely used for feature extraction and learning latent representations of data, notable issues limit their applicability to certain types of problems.[50] In general, data encoding-decoding often acts as a low-pass filter due to the model's information bottleneck failing to capture either high-frequency or fine details of data given its limited carrying capacity. This limitation often makes VAEs non-ideal for tasks in which subtle variation learning is required, as the models tend to learn axes of greatest variation first, which may compromise the ability of a model to learn distinguishing features across less significant axes. Further, because VAEs penalize the distribution of the latent features

across sampled data, representations of outlier data points are often coerced towards the median distributions of the data. The body of literature dedicated to refining and improving autoencoder designs for feature learning is actively developing, and although limitations persist, even simple models designed to learn optimal encoding functions of complex data have demonstrated capacity to extract the gestalt of a dataset across features that may be visually interpretable by a human reader, as will be shown in Chapter 2.

Relationship to Human Perception

The design of the CNN is inspired by, but not intended to mimic, the human perception system. The nature by which convolutional layers update their weights to better recognized labeled objects has drawn comparisons to the human nervous system's process of updating biological neural networks through development.[53,54] The intuition provided by the biological model of vision identifies similarities in the hierarchical arrangement of stacked convolutional layers with the processing steps involving simple and complex cells in the occipital lobe of most brains, as diagrammed in Figure 1.11. While this ideation supposes that stacks of convolutional layers add complexity and depth to representation by combining primative abstractions to represent complex imaging data, efforts to understand precisely what is represented by distinct convolutional kernels remains an on-going challenge in the field.[55] Efforts have been made to interrogate individual kernels with natural color depth to achieve a human-perceptible representation of the learned features. Figure 1.12 illustrates a few of these representations from a convolutional neural network trained to classify natural images.[56] Though interesting to look at, it remains challenging for humans to apply semantic meaning to structures or features represented by kernel weights of a convolutional

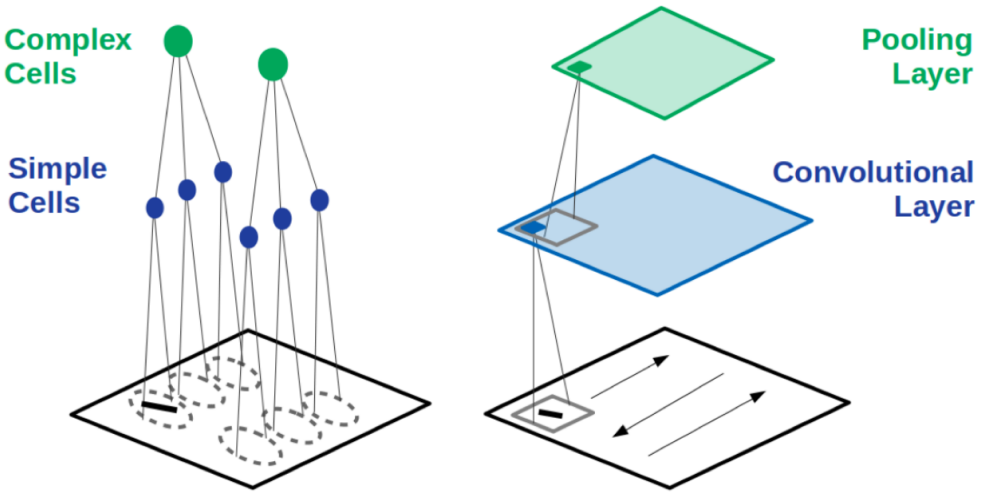


Figure 1.11: The use of convolutional operators is inspired by simplified understanding of the human perception system, which integrates stacks of interconnected neurons to execute high-level interpretation of visual stimuli.[53]

neural network, opening up fair challenges to the underlying approach as a *black box* model, though the process of dissecting the inner mechanations of any human's biological occipital processing networks remains an equally challenging prospect.[57]

1.3 Digital Pathology & Artificial Intelligence

The degree to which computers can learn relevant associations between patterns in images of patient tissue and clinical variables of interest in a manner similar to a pathoogist is evolving rapidly, and underpins most of the major recent successes at the intersection of digital pathology and artificial intelligence.[32] Although the application of deep convolutional neural networks to image analysis is undergoing renewed interest, the first documented evidence of their applciation to histopathology dates back to the early 1990s.[58,59] Even before then, the term *telepathology* was first coined in 1986, referring to the use of television

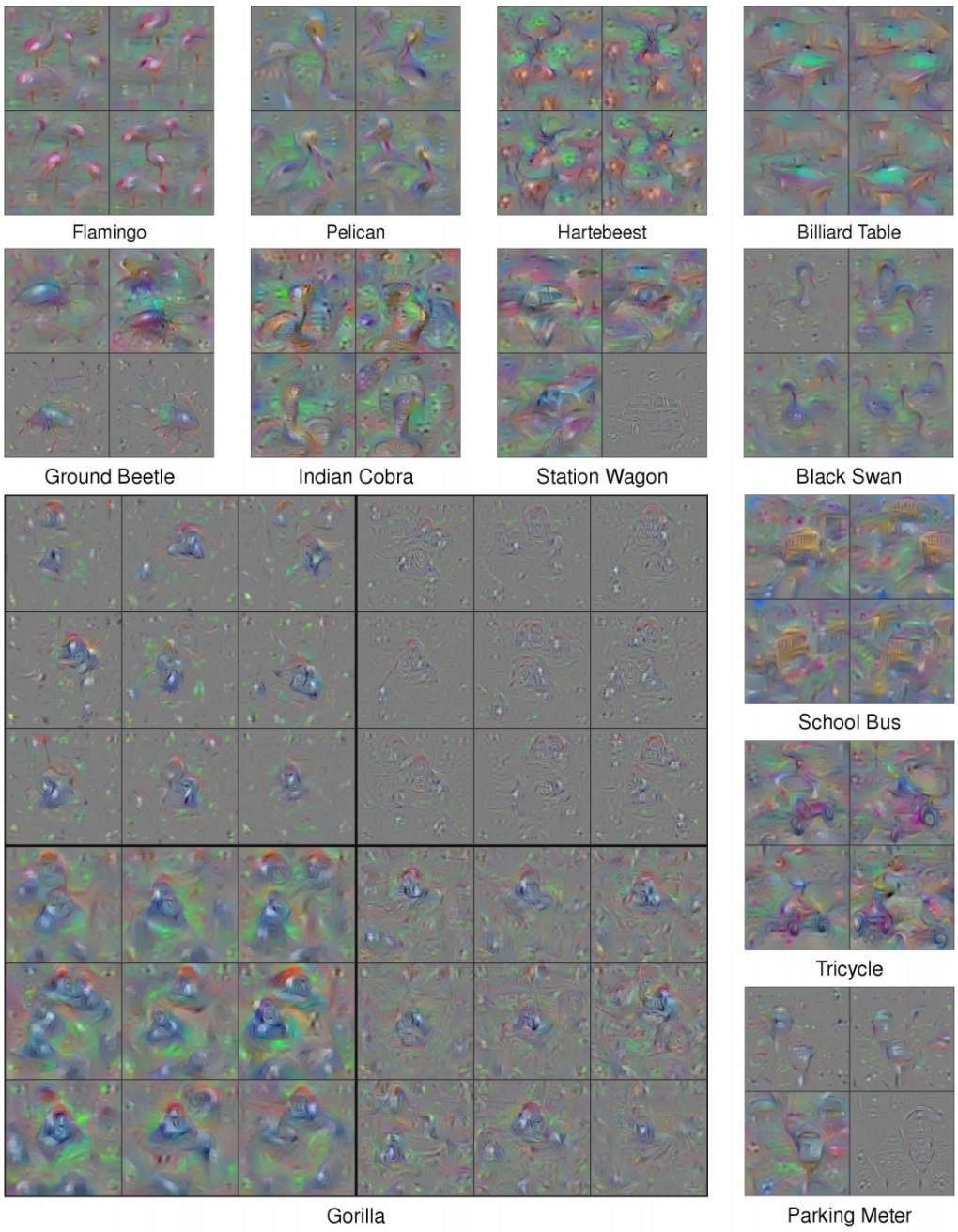


Figure 1.12: Pseudo-color examples of visualization of learned convolutional weights help intuit learned structural features neural networks use in classification tasks. Adopted from Yosinky, 2015[56]

screens to visualize tissue.[60] Since then, an array of new tools have emerged for the storage, visualization, and annotation of digitized whole slide images,[61,62] which are enabling large

and openly accessible digital slide archives such as The Cancer Genome Atlas.[63] The field continues to undergo rapid developments as the shift from analog to digital pathology continues to accelerate, particularly in light of research suggesting equivalence between digital pathology and glass slide microscopy for primary histopathological diagnosis.[64,65]

Other opportunities stemming from integrating artificial intelligence methods into medical practice are myriad.[66,67] Over the past decade, computational advances in the field of deep learning have revolutionized computer vision and digital image processing, which have been successfully applied to medical imaging in diverse fields of medical imaging including dermatology,[68] retinopathy,[69] and radiology.[70] In digital pathology, deep learning has been applied to a wide variety of tasks,[30,67,71–73] ranging from low-level operations such as object detection[74] and segmentation[75] to sophisticated clinical operation such as predicting disease prognosis[76] or response to treatment.[77] Despite the wide array of specific clinical tasks that deep learning is becoming involved in, many of fundamental steps in bringing a model to bear follow a relatively common process.[72] A graphical summary of the general process of integrating deep learning systems with pathology is shown in Figure 1.13 in which digital images are combined with clinical labels or human annotations to inform a computer vision model.

The image recognition capabilities of certain deep learning models designed to discriminate histological images across different tissue and tumor types have been shown to meet or exceed clinical benchmarks in a a number of cases.[78–80] A recent study illustrated the capacity of deep learning systems to bridge the intersection of imaging and genomics research, as a computer vision model achieved clinical grade accuracy in discriminating between lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) as well

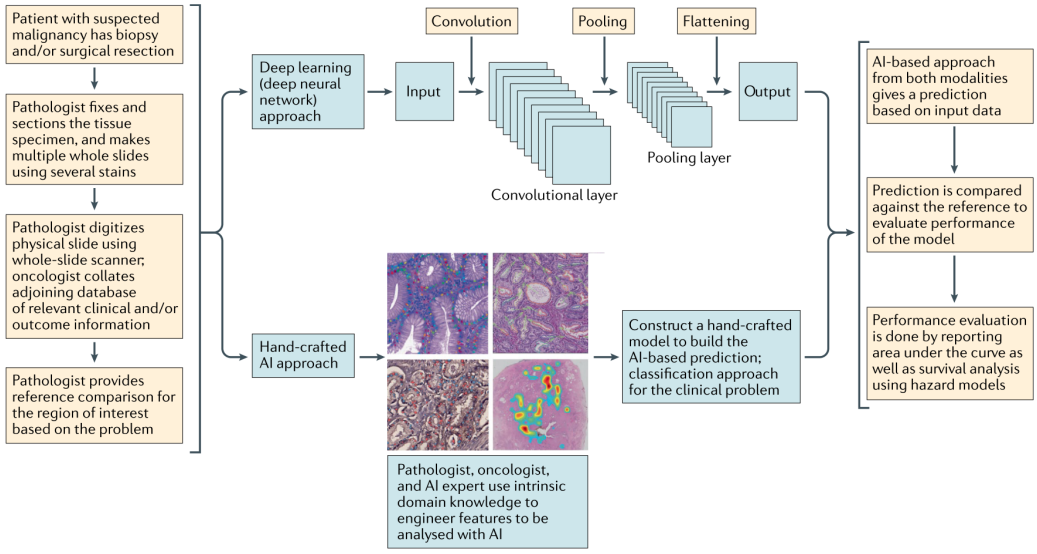


Figure 1.13: General purpose deep learning work-flow for deep learning in digital pathology. Expert guidance from pathologists in feature engineering selection controls which features in large whole slide images are relevant to the learning task.

as predict genomic mutations through histopathological spatial signatures.[81] While deep learning methods alone can demonstrate efficacy in limited settings, pathologists are trained to identify features of tissue across a wide range of clinical contexts. These and similar findings suggest ample opportunity to integrate deep learning approaches to augmented the capability of clinical pathologists.[82,83] For example, an augmented reality microscope has been shown to rapidly identify malignant features of histopathology before a pathologist inspects a slide with the intention to guiding a pathologist's attention, accelerate tissue inspection, and improve ease of diagnosis.[84]

Deep learning has also had success in addressing some of the technical artifacts associated with H&E preparation, particularly in the field of color normalization.[85] While several normalization methods[86–89] seek to address issues associated with uneven stain distribution within a slide, pathologists are expected to render diagnoses from inconsistently-treated

tissues. As a result of other technical factors including variable cutting thickness, tissue preparation, method of tissue resection or biopsy, time to fixation, operator bias, and the choice of scanner, H&E images are not consistent even within a common cancer or tissue type.

Although much of histopathology relied on H&E images, axillary staining techniques, particularly immunohistochemistry, are regularly utilized in clinical pathology. The introduction of the generative adversarial network by Ian Goodfellow and colleagues in 2014[90] has opened up novel lines of research investigating synthetically-generated virtually-stained images that are difficult to discriminate from real images. Virtual staining describes the process by which unstained tissue sections are captured digitally and a neural network applied to predict the presence of absence of particular stains of interest. A recent study demonstrated that synthetic H&E stained whole slide images made directly from an unstained bright-field image were indistinguishable from truely stained tissue sections by a set of board-certified pathologists.[91] Other recent work has employed generative models to accurately infer the presence of proteomic staining measured with immunofluorescence by learning a transfer function between whole slide H&E images and adjacent sections stained with immunofluorescently-tagged antibodies of DAPI and PanCK.[92] In general, axillary staining of specific molecular biomarkers is done via immunohistochemistry, however the choice and selection of marker panel to order can be subject to clinical bias.[93–95] Virtual staining methods may provide a computationally-efficient mechanism to infer multiplexed staining patterns from a single tissue source without the need for expensive and time-consuming companion diagnostics.

Metastatic cancer detection has attracted considerable attention in digital pathology.

Findings published in 2017 summarized top-performing models trained to identify metastatic cancer within lymph nodes, and found that top-five performing algorithms performed comperably to a set of board-certified pathologists.[76] Deep learning methods have proven to facilitate improved detection of histopathological features of metastatic breast cancer within lymph nodes, with findings supporting concordance between features detected by computer vision models and confirmatory staining done by immunohistochemistry. In research led by Martin Stumpe in 2018, a research group observed that neural network-assisted pathologists demonstrated higher accuracy than either the algorithm or the pathologist alone, enabling quantifiable synergistic outcomes resulting from engineering deep learning models into established clinical workflows.[96]

Deep learning systems have also shown to have prognostic value, as demonstrated in a recent study illustrating that neural networks in conjunction with feedback from pathologists can automatically detect the spatial organization of tumour-infiltrating lymphocytes (TILs) in images of tissue slides from The Cancer Genome Atlas and that this feature was prognostic of outcome for 13 different cancer subtypes.[97]

1.4 Challenges and Opportunities

Despite widening breadth of opportunity to address clinical needs in digital pathology with artificial intelligence, significant challenges have impeded rapid adoption[98,99] The first challenge described is that of the absence of labeled data. The adoption of the electronic health record (EHR) has in many cases provided links from digital H&E images with clinical treatments, outcomes, and diagnoses, however, lack of structure of many electronic records

presents a rate-limiting step for curating fully-labeled datasets necessary for training computer vision models.[100] Further, even if labels are found to be associated with the contends of a digitized WSI, intra-slide heterogeneity and variability often necessitate per-pixel or per-region annotations provided by the pathologist. Such overlapping or hierarchical labels of clinical interest often present non-boolean diagnostic criteria, challenging learning systems already operating on high-dimensional input data.[101]

The degree to which neural networks models are interpretable remains a significant concern and barrier to adoption of AI methods for clinical practice. A recent international study has shown that deep learning methods out-perform their human counter-parts in detecting breast cancer.,[102] though concerns have been raised regarding the transparency and reproducibility of those findings.[73,103] Open data accessibility remains another significant challenge to researchers and practitioners. The necessarily high resolution whole slide images captured by digital scanners present additional challenges stemming from the large volumes of raw imaging data, while essential protections surrounding patient privacy protection must be maintained through data-sharing mechanisms. Compared to radiology, where file sizes are typically less than 50 MB, an individual histopathological may consists of approximately tens of Gigapixels and easily eclipse several gigabytes in size.[82] The inherent size and variability within whole slide images limits the uniformity of single labels applied to an entire slide, often requiring per-pixel annotations of whole slide images to label tiles sampled from within them.[104] Although the need for large manually annotated datasets has historically posed a significant bottleneck for many deep learning applications in digital pathology, recent work led by Thomas Fuchs demonstrated that large enough datasets paired only with diagnosis labels may obviate the need for pixel-wise annotations

while retaining clinical-grade performance metrics in identifying metastatic cancer of the lymph node.[105]

Clinical medicine is often confronted with cases rare in the general population such that robust datasets covering axes of variation for a given condition might be non-existent, presenting a class-imbalance problem when training convolutional neural networks.[106] If larger datasets exist in a related context, transfer learning has been shown to be a reasonable strategy to avert issues associated with class imbalance, as has been employed with success for breast cancer prediction.[79]

Despite concerns regarding the accessibility and transparency of deep learning in digital pathology, tremendous opportunity nevertheless exists, particularly in the fields of research and education. A current decline in practicing pathologists has necessitated rethinking in terms of how pathologists are educated and whether clinical decision support systems may help alleviate work volume.[107] In cases where a pathologist might experience significant redundancy in day-to-day slide inspection, deep learning systems may be compelling mechanisms to triage difficult cases from simpler ones, providing a pathologist with more challenging cases and reducing redunant workflows for more trivial cases.[98] Aritificial intelligence systems may have a more central role to play in the education process of pathologists in the near future, as synthetic image generating systems have already been used to generate case and control samples to test a trained pathologist's capacity to discriminate between real and fake images.[71] Artificial intelligence has also shown promise in supporting quality assurance in digital images. A recent study employing deep learning to infer proteomic signature of pancreatic adenocarcinoma identified regions of tissue correctly predicted by a neural network, but whose true stain failed to stain a part of tissue due to technical issues

during the staining process.[108]

Digital pathology and artificial intelligence have, and will likely continue to offer complementary approaches to clinical practice by pursuing mechanisms by which to augment a human pathologist, identify spatial bio-markers of clinical utility, and accelerate the rate at which data can be ingested and compared.

1.5 Summary

Cancer is a pressing global health challenge, and though therapeutic advancements have and are continuing to evolve rapidly, deeper understanding of individual cancers is required to elucidate the mechanisms by which certain patients respond to therapy while others do not. Metastases present a particular threat, as diverse microenvironmental conditions of the secondary host site may differentially modulate the biology of a primary tumor or present physical barriers that impede drug diffusion. Characterizing metastatic disease is therefore a highly-relevant clinical challenge to ensure that a cancer that has metastasized is accurately diagnosed and treatment appropriately prescribed to halt further spread. The diagnostic criteria and mechanisms to do this are challenging and subject to inter- and intra-operator discrepancies, often relying on analog inspection of tissue under a microscope. While machine learning and computer vision systems are vulnerable to biases themselves, computer-aided diagnoses have demonstrated compelling potential to improve the quality of patient care by augmenting human view and understanding of medical information. The degree to which modern approaches of computer vision and machine intelligence may augment understanding of the spatial biology of tumor tissue is an area of rapid development.

This work seeks to evaluate these types of systems in the context of spatial systems biology.

1.5.1 Contributions

This thesis contributes to the study of metastatic cancer through the application of artificial intelligent systems to cellular imaging data and histopathological whole slide images. This thesis is organized into four principle chapters of original research followed by a final chapter summarizing conclusions and future research directions. Chapter 2 evaluates the role of microenvironmental perturbations on multicellular organization and demonstrates an unsupervised computer vision systems designed to learn meaningful representations of characteristic growth phenotypes. Chapter 3 introduces the problem of metastatic origin inference and presents a deep-learning system trained to characterize whole slide histologies of metastatic liver cancers by their tissue of origin. Chapter 4 introduces a novel annotation tool designed to resolve a critical bottleneck in digital pathology that incorporates unsupervised feature manifold learning to accelerate the rate at which whole slide annotations may be generated. Chapter 5 expands upon the work presented in chapter 3 by incorporating a transfer learning paradigm that integrates whole slide images of primary cancer to augment a metastatic tumor classification model. Chapter 6 summarizes this work and describes future directions for research in artificial intelligence in digital pathology by reflecting on the promises and limitations of these approaches.

Chapter 2

Multi-cellular Feature Representation Learning

> Everything we see is perspective,
> not truth.
>
> *Marcus Aurelius*

This work has been formatted for inclusion in this dissertation from the manuscript "Variational autoencoding tissue response to microenvironment perturbation', by Geoffrey F. Schau, Guillaume Thibault, Mark A. Dane, Joe W. Gray, Laura M. Heiser, and Young Hwan Chang published in the Proceedings Medical Imaging 2019: Image Processing of the International Society for Optics and Photonics in March, 2019 (doi: `https://doi.org/10.1117/12.2512660`).[109] This chapter also includes an additional section on the HCC1143 cell line that is unpublished but presented at the 19th annual International Association of Breast Cancer Researchers conference held in Egmond an Zee, The Netherlands.

Abstract

This work applies deep variational auto-encoder learning architecture to study multi-cellular growth characteristics of human mammary epithelial cells in response to diverse micro-environment perturbations. Our approach introduces a novel method of visualizing learned feature spaces of trained variational auto-encoding models that enables visualization of principal features in two dimensions. We find that unsupervised learned features more closely associate with expert annotation of cell colony organization than biologically-inspired hand-crafted features, demonstrating the utility of deep learning systems to meaningfully characterize features of multi-cellular growth characteristics in a fully unsupervised and data-driven manner.

2.1 Introduction

The presence of constituent components within the cellular microenvironment and their effect on growth, differentiation, and therapeutic response of tissue is of paramount importance in the field of spatial systems biology.[110,111] Recent advances in high-throughput systematic screening technologies enable quantification of phenotypic differences among a variety of cell populations in response to diverse chemical and genetic treatments.[112,113] The microenvironment microarray (MEMA) platform[114,115] is designed to generate images that capture diverse phenotypic changes of cellular populations exposed to soluble ligands and insoluble extracellular matrix (ECM) proteins. High-throughput generation of these types of data require powerful and sophisticated algorithms to capture features of interest to better form and validate biological hypotheses. Presently, image-based cell profiling methods utilize classical image quantification approaches to extract hundreds of features from high content images to quantify the perturbations' effects on feature gradients. However, defining and characterizing micro-environment-dependent multi-cellular spatial organization has remained an unmet computational challenge. Although popular techniques extract features such as cell counts, cellular spatial relationships (i.e., neighborhood information), or distance to cells in a specific sub-cellular structure,[113] these features are limited to characterizing spatial organization of individual cells and often require significant biological expertise to design.

In biomedical imaging analysis, deep learning techniques that employ convolutional neural networks (CNNs) to extract deep hierarchical spatial features directly from raw pixel image data have been shown to outperform classical methods that analyze hand-crafted

features.[116] Applications of deep convolutional neural networks in cellular imaging have shown promising utility for classification, segmentation, and dimensionality reduction in diverse biomedical contexts.[117,118] Multi-agent learning models, including generative adversarial networks (GANs)[90] and variational autoencoders (VAEs),[49] have recently been shown to be capable of learning salient features of high-throughput imaging screens at cellular and sub-cellular resolution.[92,116] Although powerful, GAN architecture has been shown to struggle in capturing multiple modes of input data, which limits the interpretability of their learned features.[119] Unlike GANs, VAE latent features conform to expected prior distributions, which enables elegant interpretation and visualization of what these models learn. To characterize features of multi-cellular growth patterns associated with micro-environment perturbation, this work applies convolutional variational auto-encoding architecture to analyze images of normal human mammary epithelial cells grown on the MEMA platform. Our approach confers two primary advantages. Unlike current image-based cell profiling methods that focus on single-cell analysis, our approach is designed to learn biologically meaningful spatial organization of multi-cellular populations, and we introduce a novel method to visually interpret meaningful high-dimensional learned features of a VAE model by generating synthetic samples within the principal component plane of the model's learned feature space.

2.2 Methods

2.2.1 Deep Variational Autoencoding Networks

The variational autoencoder (VAE) architecture introduced by Kingma and Welling[49] is designed to elucidate salient features of data in a data-driven and unsupervised manner. A VAE model seeks to train a pair of complementary networks: an *encoder* network θ that seeks to model an input x_i as a hidden latent representation z, and a *decoder* network ϕ that seeks to reconstitute x_i from its latent representation z. The VAE loss function shown in Equation 2.1 regularizes model training with an additional Kullback-Leibler (KL) divergence term that penalizes the distribution of z with respect to a given prior, which in our case is the standard normal Gaussian distribution, $p(z) = \mathcal{N}(0, 1)$. By specifying a latent dimension z less than the input dimension of x_i, a VAE model learns optimized encoding and decoding functions that enable reconstruction of an input sample subject to capacity constrains of the latent feature space within the model.

$$\mathcal{L}_i(x_i, \theta, \phi) = -\mathbb{E}_{z\sim q_\theta(z|x_i)}[\log p_\phi(x_i|z)] + \mathrm{KL}(q_\theta(z|x_i)\|p(z)) \tag{2.1}$$

The VAE model trained in this study incorporates two-dimensional convolutional layers to encode spatial information of multi-cellular organization of cells grown in diverse micro-environments. Specifying a limiting bottleneck on the latent feature space forces the model to learn salient features of the dataset and reduce the dimensionality of input features for further downstream analyses.

Learning Model Design

The encoder and decoder models used in this study are congruent and composed of five 2D convolutional layers each containing 64 filters with same padding and rectified linear unit activations on all layers except for the final sigmoidal decoder layer. The outer two convolutional layers have a 3×3 kernel, the inner two layers have a 2×2 kernel, and the latent layer is composed of 16 hidden features, which illustrated good trade-off between model capacity and training loss. Both the encoder and decoder are optimized with the RMSProp optimizer against the custom variational loss function that penalizes the binary cross entropy between input and reconstruction as well as the KL divergence between the latent space sample and standard normal distribution. The models designed for this study were written in Python using Keras[120] with Google's Tensorflow backend,[121] and trained using Nvidia Tesla V100 GPUs mounted on the Exacloud high performance computing environment at OHSU.

2.2.2 MEMA Dataset

This study seeks to uncover the role of microenvironmental perturbations in the growth of normal human mammary epithelial cells (HMECs) by evaluating phenotypic response to 57 ligands and 47 extracellular matrix (ECM) components using the microenvironment microarray platform,[114] reproduced in Figure 2.1. In this assay, ECM proteins are robotically printed into micro-well plates to form 300 μm spots upon which cells bind and grow. Additionally, soluble ligands are added to each well, thereby creating a combinatorial microenvironmental perturbation comprised of one ECM and one ligand per spot. After three days of growth, cells are fixed and stained for Keratin 19 (luminal marker in the red chan-

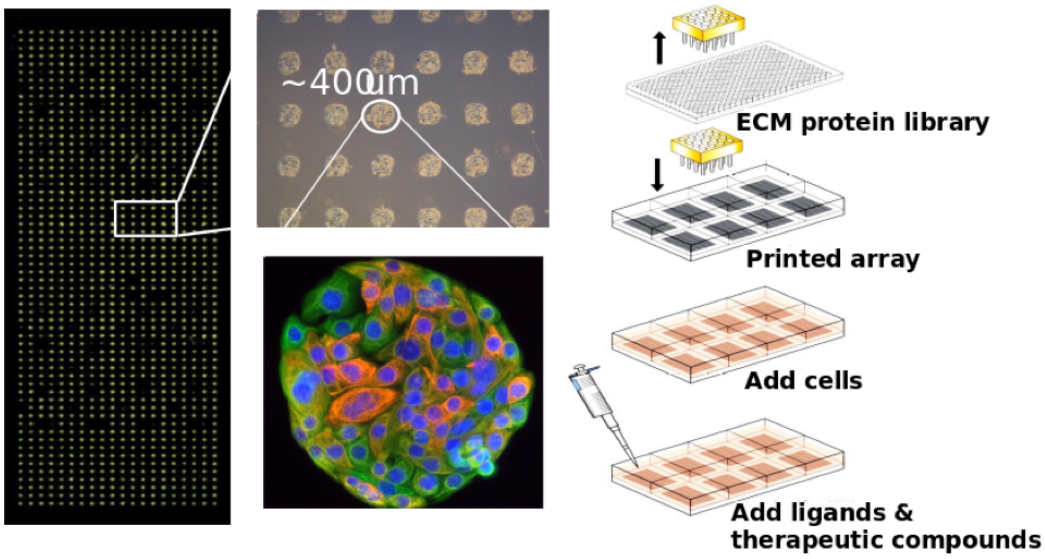


Figure 2.1: The Micro-environment Micro-array (MEMA) platform is designed to treat and fix small cellular populations on discrete spots in which each spot if treated with unique combinations of extracellular matrix component and ligand.

nel), Keratin 5 (basal marker in the green channel), and DAPI (nuclear marker in the blue channel). Input data from this study are 37,269 images of individual MEMA spots down-sampled from full-resolution (1200×1200) to 256×256 pixels. Detailed experimental description, data, and meta-data of the data-generating process are available at the MEP-LINCS Synapse wiki.[1]

2.3 Results

These results present both published and unpublished findings for both normal human mammary epithelial cells (HMECs) and HCC1143 triple negative breast cancer.

[1] `https://www.synapse.org/mep_lincs`

2.3.1 VAE Analysis

A VAE model was trained for 100 epochs on the 37,269 MEMA spot images evaluated in this study. Input image reconstructions shown in Figure 2.2 illustrate that the trained model learns sufficient spatial features of spot organization to reconstruct an input image from 16 learned latent features. Although the reconstructions are clearly lossy, they suggest that organization, intensity, and distribution of signal within the spots is learned. Notably, despite the clear heterogeneity in the dataset, the learned reconstructions are generated from a set of 16 learned features that conform to the expected standard normal prior placed on the learning loss function. Because the prior places no constraint on relationships between learned features, correlations between learned features exist. Interestingly, both the number of cells on each spot and localized abundance of the KRT19 luminal marker, both of which are typically used to characterize spot organization, appear to associate within the learned VAE feature space, which is visualized in two dimensions with the t-SNE algorithm.[122]

Local sub-regions of the learned VAE feature space are further visualized in the two-dimensional t-SNE projection by superimposing the input images onto the t-SNE coordinates as illustrated in Figure 2.3. By examining the embedding space in this manner, local regions of the learned feature space appear to group MEMA spots by similar features such as shape, color, and morphology.

2.3.2 Latent Space Walking

To provide a qualitative assessment of the learned VAE features, we employ a latent space walking procedure that holds all but one learned feature fixed at the latent dimension's expected value (zero) while the feature of interest is swept through the inverse cumulative

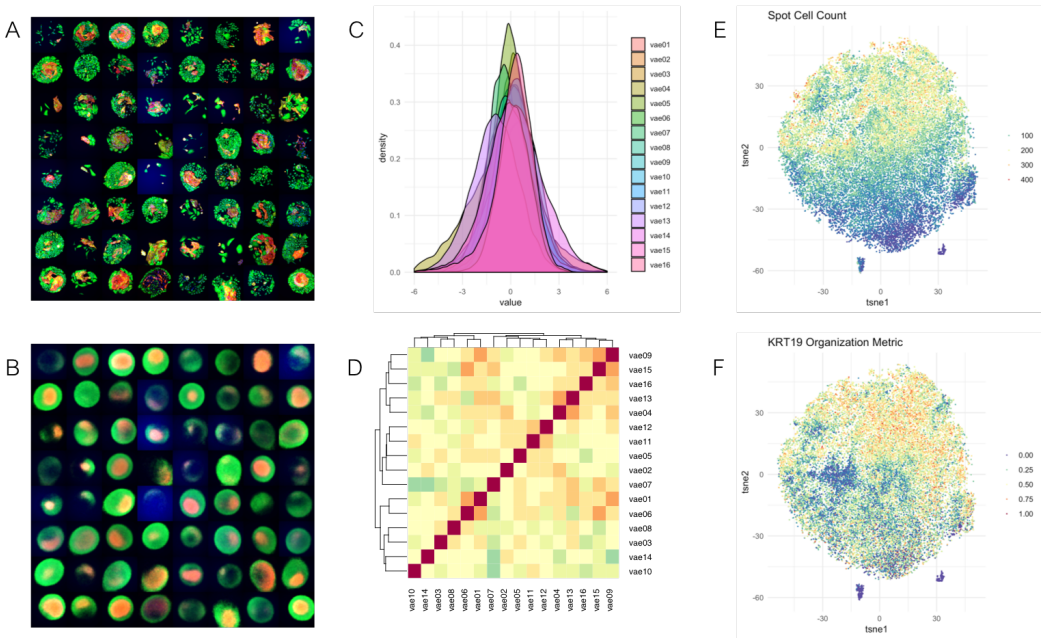


Figure 2.2: (**A**) randomly sampled input images from the full dataset (**B**) lossy reconstructions of the sampled input images after training (**C**) distributions of each of the 16 features across the entire dataset (**D**) correlation heatmap of the learned VAE features (**E**) t-SNE projection of VAE space colored by the number of cells on each spot (**F**) t-SNE projection of VAE space colored by a hand-crafted feature designed to evaluate cell spot organization.

distribution function (CDF) of the standard normal Gaussian, as in Kingma, *et al.*[49] By passing synthetic latent feature samples through the trained decoder network, the VAE generates samples that correspond to changes in a single feature of interest while holding the rest constant. At left in Figure 2.4 illustrates the effect each learned VAE feature (shown in columns) has on the decoded synthetic sample by sweeping it through the cumulative density function of the standard normal distribution (shown in rows). Although this representation can provide a qualitative assessment of each of the independent learned features, this established analysis does not consider recurring correlations between independent features. The nature of neural computing and the covariance matrix shown in Figure 2.2 suggest that learned features interact in complex, non-linear ways that cannot be visualized

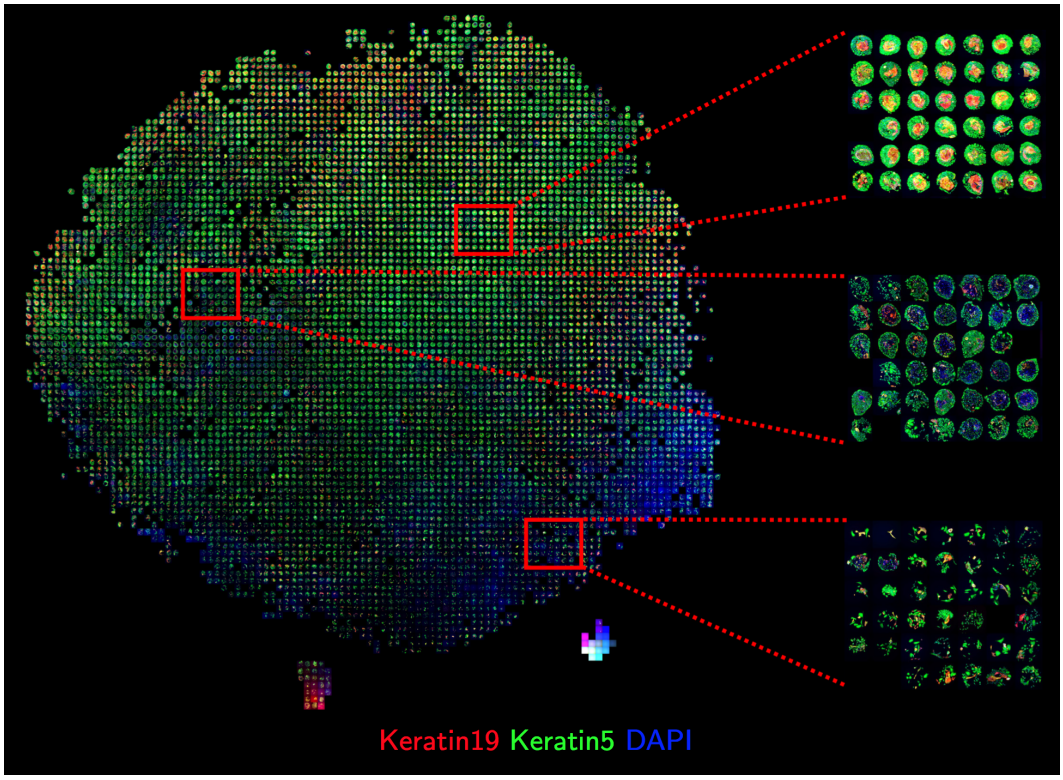


Figure 2.3: t-SNE embedding of MEMA spots used in this study based on learned latent features illustrates distinct sub-regions of feature space populated by spots of similar morphology, including a set of technical errors in the bottom right (best viewed at full 10k digital resolution).

with this class of latent space walking techniques.

To improve the degree to which the learned model's feature space may be interpreted, we introduce a novel principal feature manifold (PFM) visualization approach. Our technique is based off principal component analysis (PCA), which computes a set of principal components that capture sources of significant variation within a dataset. In brief, PCA transforms an input dataset into projection matrix $\mathbf{T}$ by rotating the input data $\mathbf{X}$ by a computed weight matrix $\mathbf{W}$, which is derived from the eigen decomposition of the data's covariance matrix, such that $\mathbf{T} = \mathbf{XW}$.

To do so, we first reduce the learned VAE feature space to the first two principal com-

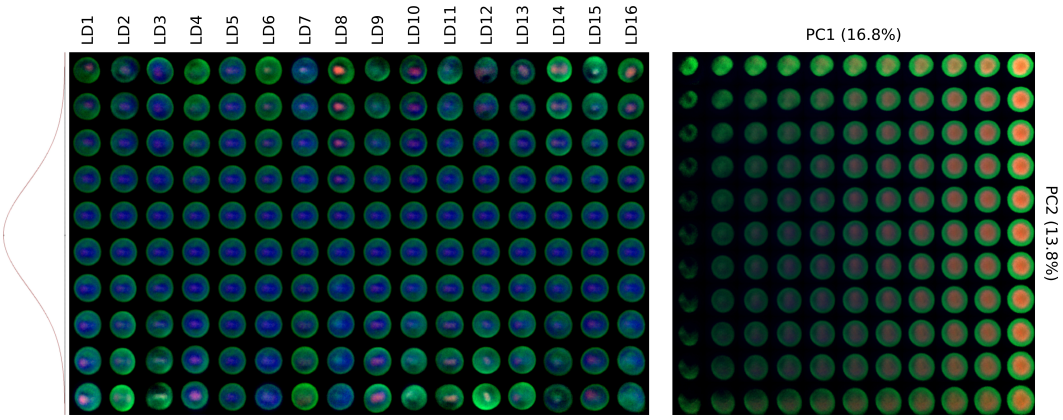


Figure 2.4: (left) Latent space walking where each column represents one of 16 latent variables in the VAE model and each row represents uniformly spaced samples along the cumulative density function of the latent variable distribution. (right) The principal feature manifold sampled from the first two principal components of the learned VAE feature space embedding visualizes sources of significant variation in the VAE encoding dataset in two dimensions. In this analysis, the first two components explain 16.8% and 13.8% variance, respectively.

ponents using PCA. Because the variance of each principal component is known, we then sample a bivariate percentile distribution $\hat{\mathbf{X}}$ that is scaled to the variance of the first and second principal components to span the sampling space we wish to visualize. We next multiply the sampled percentile grid by the inverse of the principal component matrix $\mathbf{W}^{-1}$ to rotate the uniform grid back into VAE feature space. The trained decoder network ϕ then transforms the resulting VAE space samples into synthetic input images $\mathbf{S}$ which can be visualized in two dimensions, as shown at right in Figure 2.4.

$$\mathbf{S} = \phi(\hat{\mathbf{X}}) = \phi(\mathbf{T}\mathbf{W}^{-1}) \tag{2.2}$$

This approach illustrates variability in the learned feature set by decoding higher-dimensional feature interactions presented in the first principal component plane. Although the information contained in a classical latent space walk and the introduced principal fea-

ture manifold are similar, the PCA formulation enables evaluation of the entire latent feature space in a simple two-dimensional image. While similar to the t-SNE space embedding shown in Figure 2.3, the PFM approach is uniquely capable of generating arbitrary synthetic samples using the trained decoder model.

2.3.3 Measuring Organization with Human Annotation

Presently, measuring organization of MEMA spots requires single-cell segmentation and feature extraction to first classify every cell on the spot as either basal or luminal based on expression of keratin markers. Spot organization is then computed as hand-crafted metric that measures relative abundance of keratin 19 (KRT19), a structural component of epithelial cells, within the central core region of the spot with respect to the outer region. Although reasonably effective in this experiment, similar types of hand-crafted features require sophisticated pre-processing steps and special knowledge of the biological phenomena under study to design effectively which profoundly limits translation of one such metric to other problems or experiments.

To evaluate how well our model characterizes spatial organization of cells in an unsupervised manner, we incorporate annotations from 7 expert biologists who graded 300 randomly selected MEMA spots as either *unorganized* (intermixed or single cell-type populations), *partially organized*, or *well organized* (centrally clustered luminal cells surrounded by basal cells). The inter-rater agreement is measured using the Fleiss kappa metric ($\kappa = 0.473$), which suggests moderate agreement between raters[123] while reflecting the inherently subjective challenge of characterizing multi-cellular organization. Downstream analyses assign the mode rating across all raters to each of the 300 scored spots as a simple majority vote

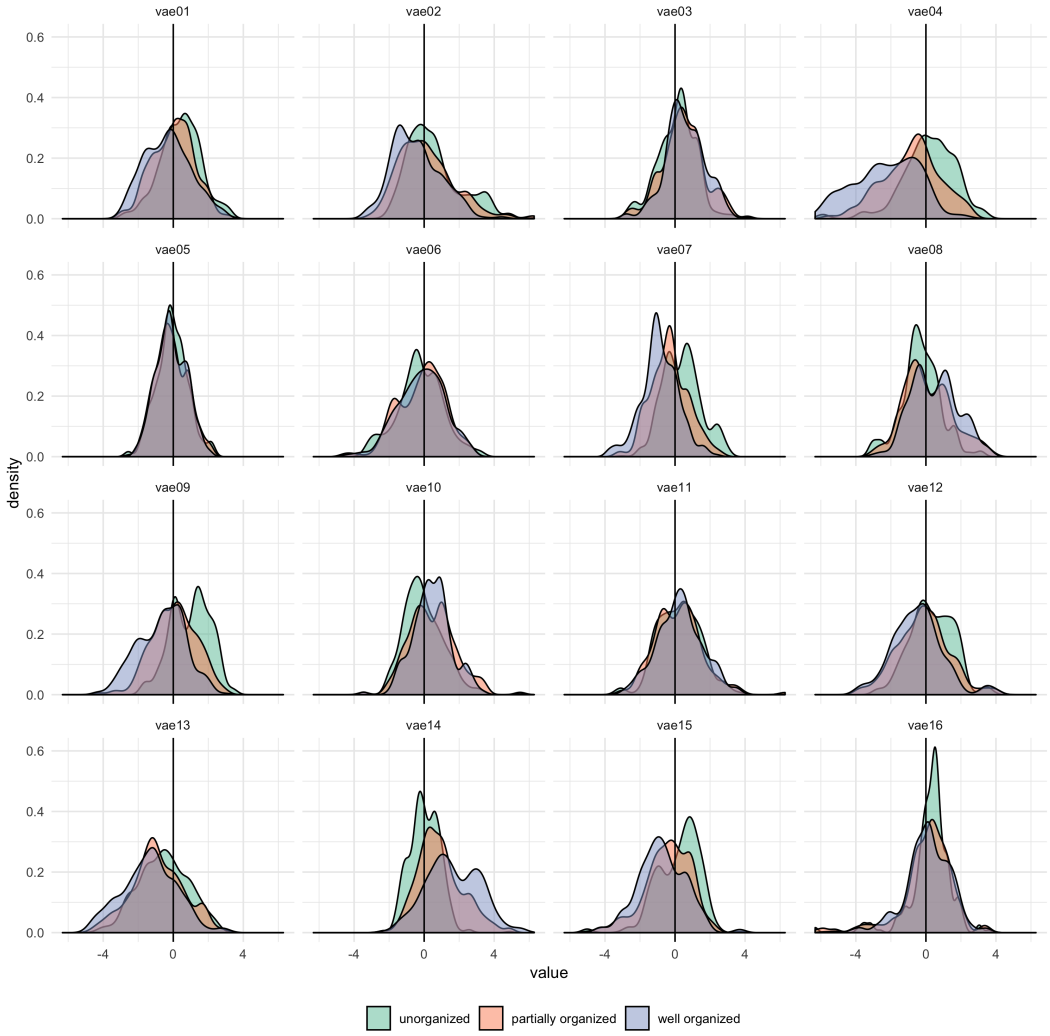


Figure 2.5: Density separation of human annotation for 300 images across 16 learned latent features.

decision. Associations between learned VAE features and annotated organization are illustrated in Figure 2.5, which suggests that certain features (particularly features 4, 7, 9, and 14) appear to exhibit shifts in their distribution with respect to organizational annotation.

To provide a fair comparison between the learned VAE space and the hand-crafted feature, we first reduce the sixteen learned VAE features into a single feature for comparison for both supervised and unsupervised settings. In this analysis, we used the first principal

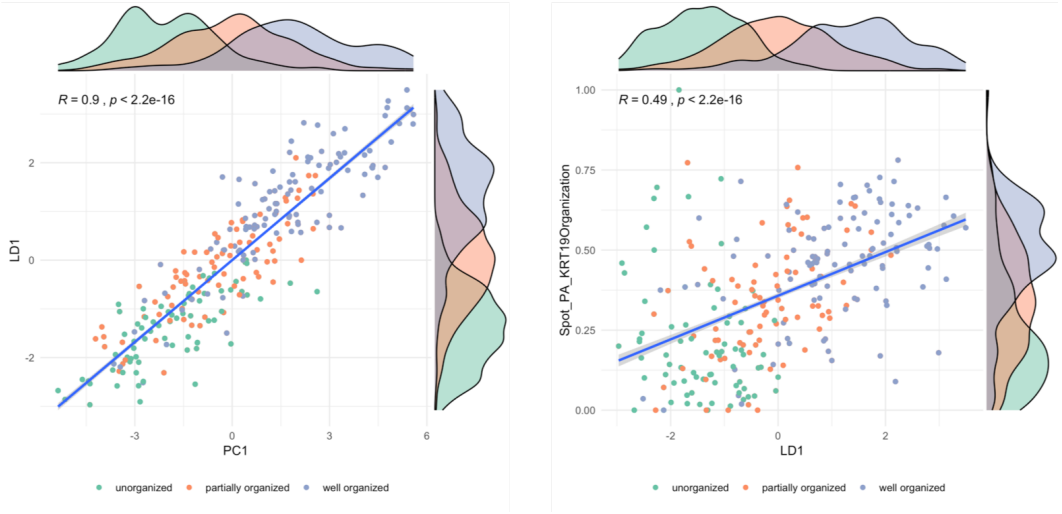


Figure 2.6: (left) The first principal component (PC1) and first linear discriminant (LD1) of the latent space are tightly correlated and illustrate clear separation between annotation class. (right) Associations between the first principal component of the VAE feature space and the hand-crafted organizational feature are weak. However, ANOVA analysis suggests that the learned VAE space improves discriminatory power between the three annotated classes.

component (PC1) for unsupervised comparison and the first linear discriminant (LD1) for supervised comparison. The relationship between the first principal component (PCA) and first linear discriminant (LDA) of the VAE latent space, shown at left in Figure 2.6, illustrates that the fully unsupervised and supervised metrics are strongly associated (Pearson correlation $R = 0.9$) while neither the first linear discriminant nor principal component correlate particularly strongly with the hand-crafted organizational feature (Pearson correlation $R = 0.49$ and $R = 0.41$, respectively). However, clear class separability is evident for both the hand-crafted feature as well as the fully unsupervised characterization by the learned VAE space, shown at right in Figure 2.6.

To test the significance of these observations, ANOVA tests compute statistically-significant separation of the three expert annotation classes (unorganized, partially organized, well-organized) with respect to the first principal components, first linear discriminant, the

Table 2.1: ANOVA results of class separation (VAE vs. hand-crafted feature (HCF) and Cell Count; *significant)

	VAE				
	PC1	PC2	LD1	HCF	Cell Count
F value	717.9	8.2	1073	254	431.5
Pr (>F)	<2e-16*	2.83e-4*	<2e-16*	<2e-16*	<2e-16*

hand-crafted organization feature, and the spot cell count. The resulting F values and associated p-values tabulated in Table 2.1 suggest that a fully unsupervised trained VAE model (F value = 717.9) improves class separability over a classically designed hand-crafted feature (F value = 254).

2.3.4 Characterizing Micro-environment Perturbation

This study was designed to evaluate the effect micro-environment perturbations (MEPs) have on cellular growth characteristics. If certain groups of MEPs (either ligands or extracellular matrix components) induce similar changes in growth morphology on the MEMA spots, and if the VAE feature space learns to capture those organizational characteristics, then similarly treated spots should be closely associated in the learned VAE feature space. This analysis first computes the mean principal latent space projection of spots treated with the same ligand-ECM combination and then performs hierarchical clustering on both ligand and ECM conditions which are shown as a heatmap in Figure 2.7. In addition to reflecting understanding that ligands have an overall more pronounced effect on cell spot organization than ECMs, this analysis highlights micro-environmental factors most strongly associated with multi-cellular organization characteristics. For example, this visualization associates certain ligands known to be highly associated with cellular growth and organi-

zation (TGFB, FGF2, FGF6, WNT3A, WNT10A, IL6, IL13, and BMP2). Interestingly, TGFβ and BMP ligands–two closely related signaling molecules–tend to be associated with cellular organization in cancer.[124] They are also implicated in epithelial–mesenchymal transition, which is relevant to the shift in KRT markers. This observation is intriguing, as these molecules are also known to play a key role in cellular differentiation and morphogenesis. Additionally, this analysis also identifies independently observed technical artifacts in a few of the ECM conditions, which are clearly distinct as the furthest two right columns of the heat map and shown as technical artifacts in Figure 2.3. Though preliminary, this type of analysis provides a rapid, unsupervised inference approach to evaluating sets of microenvironment perturbations that similarly affect cellular organization and prioritize factors for more detailed experimental studies.

2.3.5 HCC1143

HCC1143 triple negative breast cancer (TNBC) cells are grown on COL1 spots and treated with one of 63 soluble ligand conditions. Cells are grown for three days, fixed, and stained with KRT14 (luminal), VIM (mesenchymal), and DAPI (nucleus). Figure 2.8 illustrates distinct regions of the VAE encoding space showing patterns of varying degrees of cellular organization. Examining the VAE feature space with respect to individual ligand treatments illuminates which treatments tend to co-locate, as well as the consistency of features of spots undergoing similar ligand exposure across multiple technical replicates. Two dimensional density gradient lines illustrate the degree to which spots under certain treatments tend to pile up in the feature manifold in which more dense distribution of spots suggests less heterogeneity in their learned phenotypic features, while broader distribution

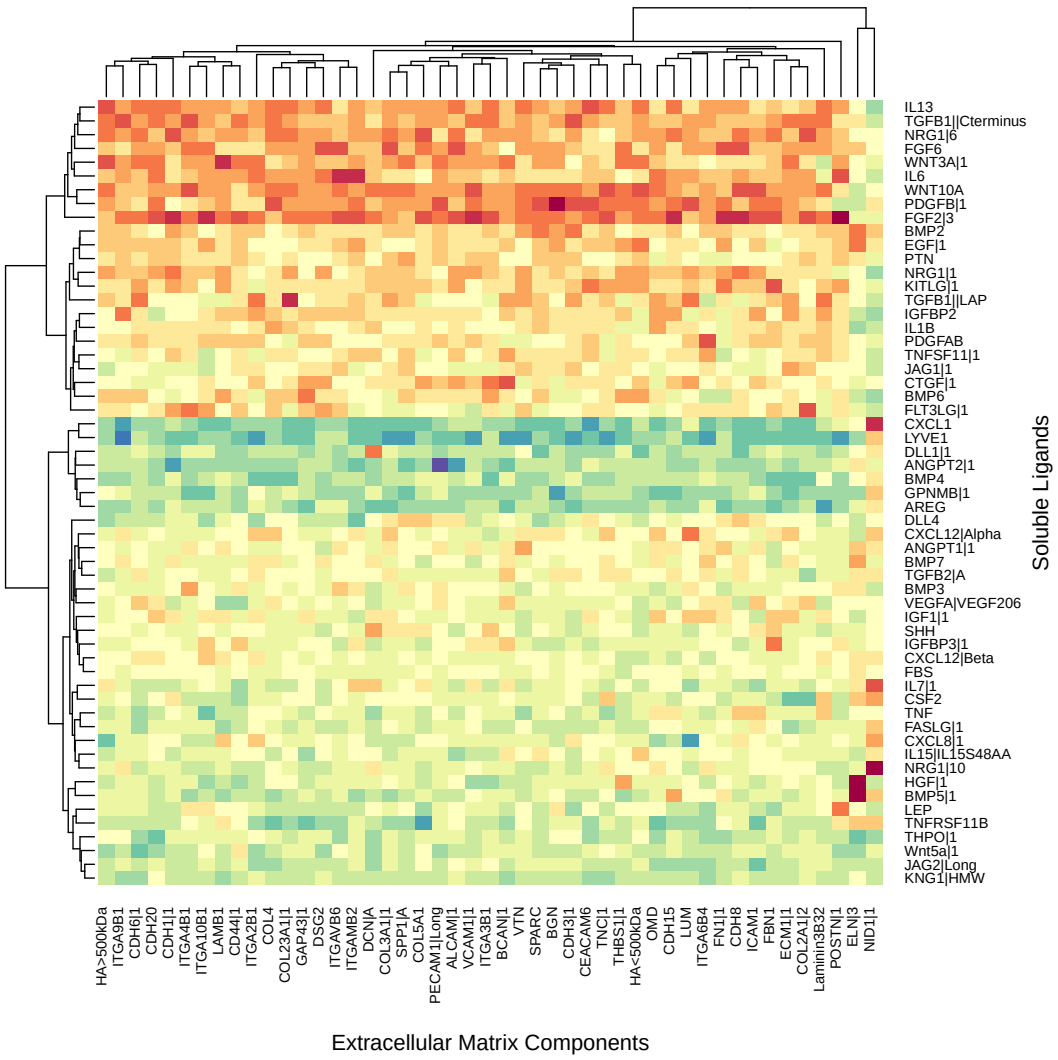


Figure 2.7: Hierarchically clustered MEPs by the mean encoding of their treated MEMA spot images by extracellular matrix (x-axis) and soluble ligand (y-axis) conditions. Each square represents the mean projection of encoded images for given ligand-ECM conditions onto the first principal component of the VAE feature space. In this illustration, red colors are more highly associated with cell spot organization and blue colors are more highly associated with cell spot disorganization (see Figure 2.4).

suggests greater variability in treatment response. Figure 2.9 illustrate which ligands tend to induce similar phenotypic responses through hierarchical clustering in which joint group membership between unsupervised clusters and groups of receptor class is visualized by a circle plot. In this approach, ligands were grouped into six clusters based on their target

class of either WNT, TNFR, TGFβ, RTK, Notch, or Cytokine and k-means clustering was performed on the latent representation with $k = 6$ to group spots based on learned morphological features, which are visualized in two dimensions in Figure 2.8. These results suggest inconsistent cross-clusterings between these two grouping methods, though some structure is observed. Ligand treatments that target cytosine-related biological function tend to impart phenotypes in cluster 1, which ligands targetting RTK (receptor tyrosine kinase) signaling appear split between clusters 1 and 4. As mentioned, these results are preliminary, but illustrate that unsupervised clustering of spots based on image feature extraction may provide avenues to draw associations between ligand target and induced multicellular phenotype.

2.4 Discussion

This work evaluates the role of variational auto-encoding models to learn latent space representations of high-throughput imaging screens of human mammary epithelial cells in response to micro-environment perturbations. We illustrate that convolutional VAE architecture provides a powerful approach for capturing high-level features that associate with expert human annotation and hand-crafted features designed to measure cellular organization. Additionally, we introduce the Principal Feature Manifold technique designed to visualize interactions between learned VAE features beyond typical latent space walking. These analyses represent a preliminary exploration into the utility of deep learning systems to capture experimentally meaningful features of spatial organization with which to characterize tissue growth patterns in response to micro-environment perturbation.

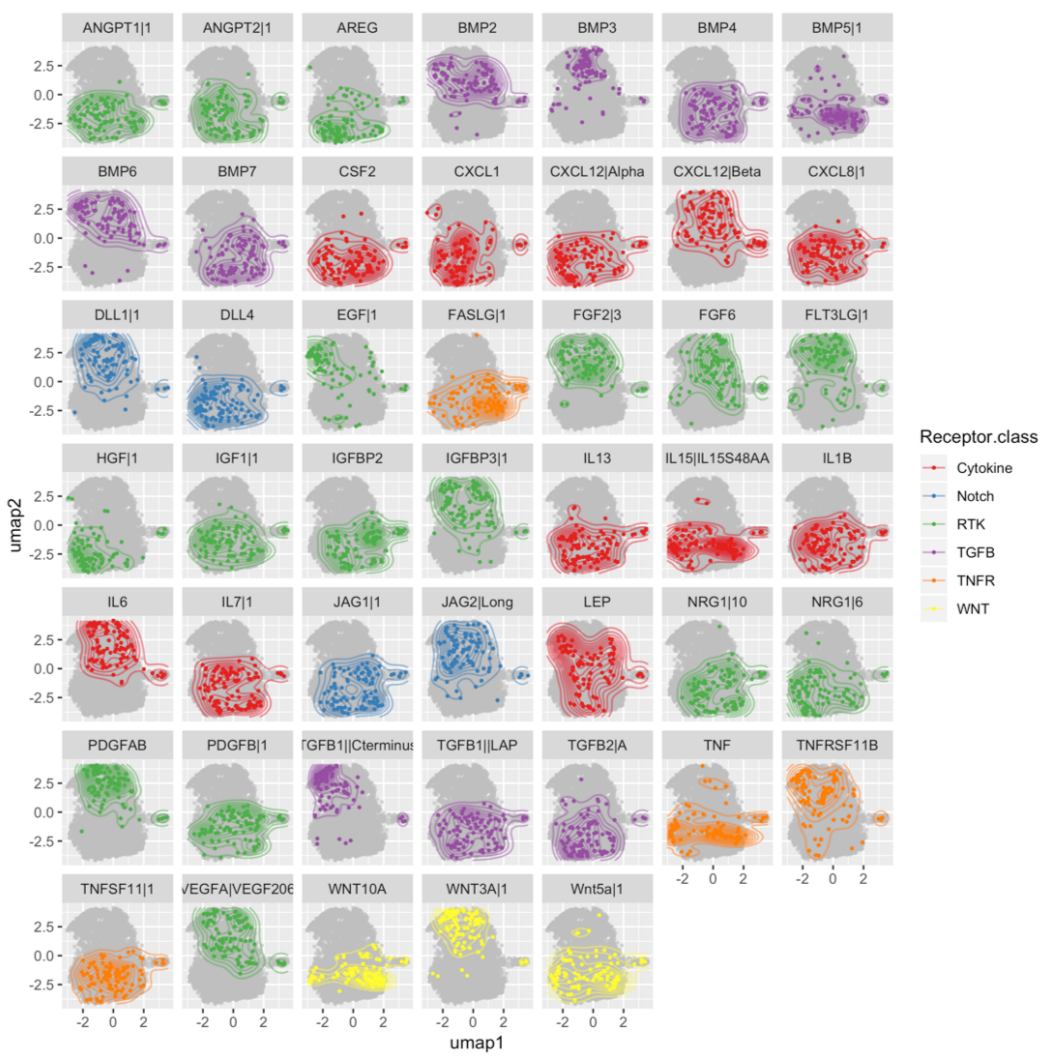


Figure 2.8: VAE embedding projection of the HCC1143 cell line perturbed by the MEMA platform

Acknowledgements

We thank Elliot Gray and Erik Burlingame for their helpful comments and discussion. The resources of the Exacloud high performance computing environment developed jointly by OHSU and Intel and the technical support of the OHSU Advanced Computing Center are gratefully acknowledged. This work was was supported in part by the NIH Common Fund Library of Network Cellular Signatures grant HG008100, the NCI U54CA209988, and the

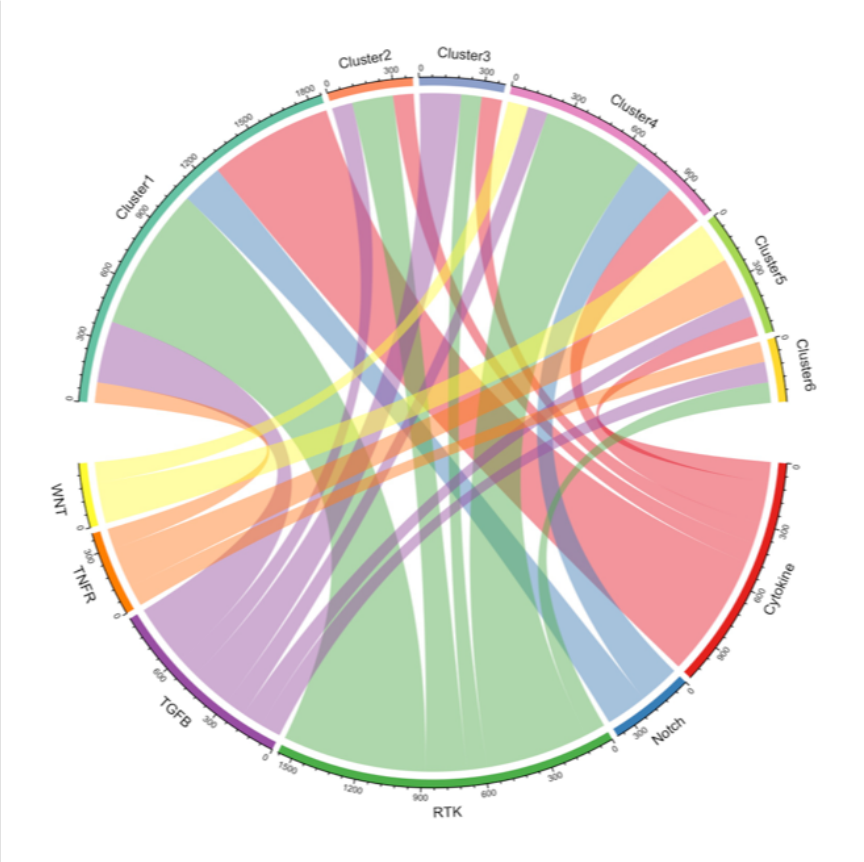


Figure 2.9: Circle plot illustrating mutual grouping of ligands by receptor class and unsupervised hierarchical clustering.

OHSU Center for Spatial Systems Biomedicine. YHC acknowledges grant support from the

Brendon-Colson Center for Pancreatic Care and CRUK-OHSU Spark Award.

Chapter 3

Neural Estimation of Metastatic Origin

I have traveled the length and breadth
of this country and talked with the
best people, and I can assure you that
data processing is a fad that won't last
out the year.

Editor of Business Books
Prentice Hall (1957)

Computers are useless;
they can only give you answers.

Pablo Picasso

This work has been formatted for inclusion in this dissertation from the manuscript "Predicting primary site of secondary liver cancer with a neural estimator of metastatic origin", by Geoffrey F. Schau, Erik A. Burlingame, Guillaume Thibault, Tauangtham Anekpuritanang, Ying Wang, Joe W. Gray, Christopher Corless, and Young Hwan Chang. This work was published in the Journal of Medical Imaging in February, 2020 (doi: `https://doi.org/10.1117/1.JMI.7.1.012706`).[125]

Abstract

Pathologists rely on relevant clinical information, visual inspection of stained tissue slide morphology, and sophisticated molecular diagnostics to accurately infer the biological origin of secondary metastatic cancer. While highly effective, this process is expensive in terms of time and clinical resources. This work seeks to develop and evaluate a computer vision system designed to reasonably infer metastatic origin of secondary liver cancer directly from digitized histopathological whole slide images of liver biopsy. This work illustrates a two-stage deep learning approach to accomplish this task. We first train a model to identify spatially-localized regions of cancerous tumor within digitized hematoxylin and eosin (H&E) stained tissue sections of secondary liver cancer based on a pathologist's annotation of several whole slide images. Then, a second model is trained to generate predictions of the cancers' metastatic origin belonging to one of three distinct clinically-relevant classes as confirmed by immunohistochemistry. Our approach achieves a classification accuracy of 90.2% in determining metastatic origin of whole slide images from a held-out test set, which compares favorably to an established clinical benchmark by three board-certified pathologists whose accuracies ranged from 90.2% to 94.1% on the same prediction task. This work illustrates the potential impact of deep learning systems to leverage morphological and structural features of H&E stained tissue sections to guide pathological and clinical determination of the metastatic origin of secondary liver cancers.

3.1 Introduction

Metastatic liver cancer accounts for 25% of all metastases to solid organs, yet because liver metastases can arise from almost anywhere in the body, accurately determining the origin of metastatic liver cancer is of paramount importance for guiding effective treatment.[126,127] In clinical practice, pathologists commonly rely on clinical information, tissue examination, and molecular assays to determine the metastatic origin of a patient's secondary liver tumor. Although clinically effective, this approach requires significant expertise, experience, and time to perform properly. Deep learning methods have rapidly accelerated the automation of key processes in identifying and quantifying clinically meaningful features in biomedical images and continue to drive modern advancements in digital pathology.[128,129] Furthermore, deep learning systems have been applied to settings where their performance matches and even exceeds the ability of clinical human practitioners in tasks related to image analysis, including in clinical instances that rely on inspection of hematoxylin and eosin (H&E) stained tissue.[74,81,130–132] The emerging power and success of many deep learning approaches applied to image content analysis stem from their ability to learn and leverage meaningful features from large data collections that cannot be explicitly mathematically modeled.[131,133–137] For example, these approaches can provide robust and reproducible solutions for automated detection and analysis of tumor lesions within whole slide images containing both normal and cancerous tumor tissue segments.[138–141]

Our key contribution in this paper is a deep learning approach to identify metastatic tissue within whole slide section and classify these tumors by their metastatic origin. We evaluate model performance with respect to a clinical benchmark established by three board-

certified pathologists charged with the same classification task as our model in which each pathologist was tasked to infer the metastatic origin of liver cancer directly from H&E stained tissue sections without the use of molecular immunohistochemistry assays or clinical data. Through this work, we demonstrate feasibility of deep learning systems to automatically characterize the biological origin of metastatic cancers by their morphological features presented in H&E tissue sections.

3.2 Methods

3.2.1 Data Set

This study collected 257 whole slide scanned H&E stained images of metastatic liver cancer. Raw H&E images were acquired from the OHSU Knight BioLibrary, uploaded to a secure instance of an OMERO server,[142] programmatically accessed through the OpenSlide python API,[143] normalized with established methods to overcome known inconsistencies in the H&E staining process,[86] and tiled into non-overlapping patches of 299×299 pixels necessary to accommodate the utilized deep learning architecture. Tiles whose mean three-channel, 8-bit intensities were greater than 240 were filtered out as white non-informative background. The total training data set is composed of twenty thousand non-overlapping tiles from tumor tissue within the H&E scanned images. Each image in the dataset is annotated with clinically-determined metastatic origin labels informed by clinical information, pathological inspection of tissue sample, and staining by immunohistochemistry (IHC). Clinical annotations were summarized into 3 distinct subgroups by a clinical practitioner, which are summarized in Fig. 3.1. Annotations of tumor regions within 28 whole-slide H&E

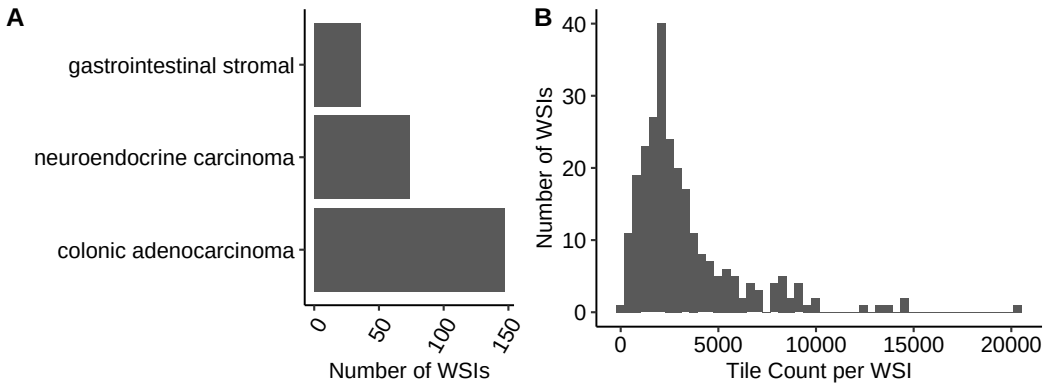


Figure 3.1: (A) Summary of the acquired dataset, composed of WSIs each containing metastatic tissue originating from one of 3 sites of interest. (B) Distribution of non-overlapping tile counts in each WSI with mean count 3300 tiles per WSI.

images were generated by a board-certified pathologist and collected using PathViewer[1], an interactive utility for the collection and storage of pathological annotations.

3.2.2 Learning Approach

Our approach is composed of two deep neural networks that operate in series. The first stage model is trained to filter tiles containing normal or stromal tissue from whole slide images (WSIs), as these tiles are not expected to have predictive value in estimating the metastatic origin of cancerous tissue. The first stage model is trained to pass through tiles containing cancerous tissue from whole slide images (WSIs) and filter out tiles containing normal liver. A second stage model is then trained to predict a single label of metastatic origin for each tile in the dataset. Individual per-tile predictions are then aggregated within their respective whole slide image and averaged to compute a single prediction for the whole slide image. A diagram illustrating the basic work-flow of our approach is shown in Fig. 3.2.

In the first stage of our approach, pathologist annotations of tumor regions are employed

[1] `https://glencoesoftware.com/products/pathviewer/`

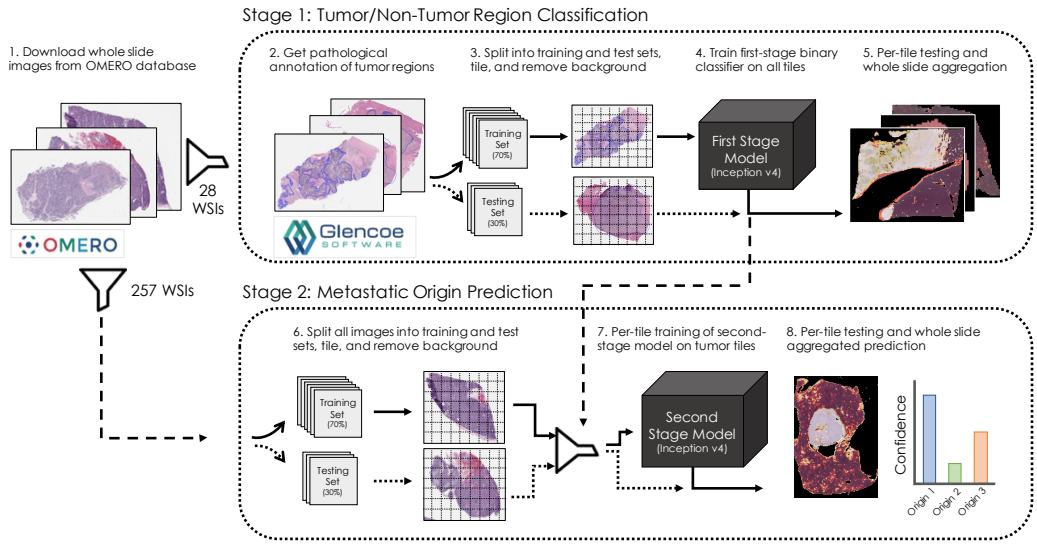
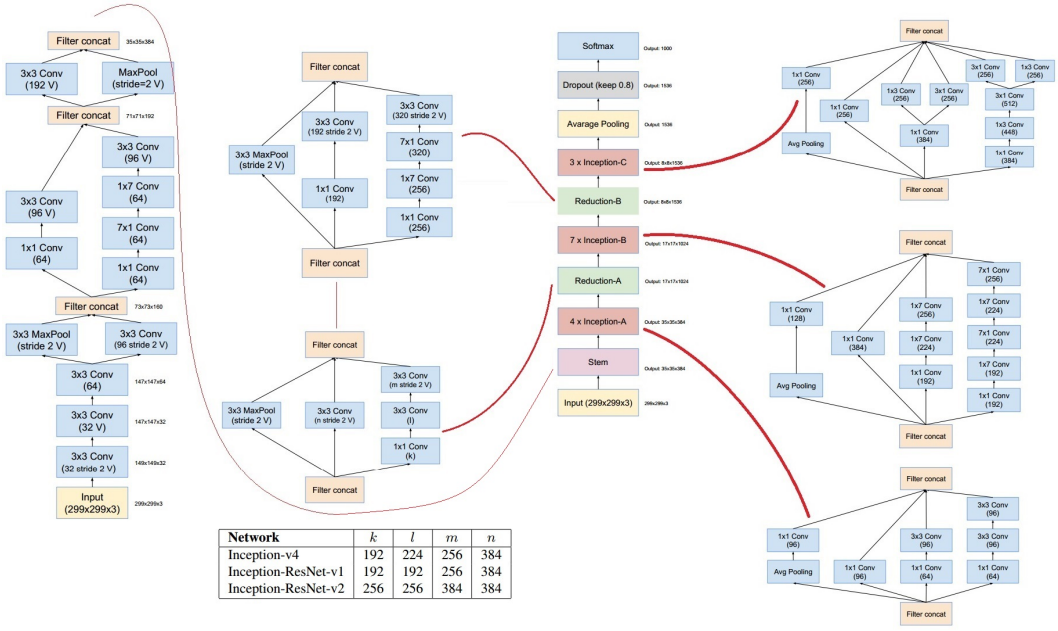


Figure 3.2: Deep learning based approach to leverage pathological annotation of tumor region to isolate and localize tumor tissue from a WSI and generate predictions of metastatic origin.

to train a binary classifier to predict whether a given tile of H&E image is either tumor or non-tumor tissue. A second stage classification model is then trained on just the tumor portions of images to predict metastatic origin based on clinically-determined whole-slide labels. In all cases, the predictions from the models are reassembled into probabilistic heat maps over the WSI, enabling a rapid assessment of spatial characteristics driving predictive reasoning. Both first and second stage models utilize the Inception v4 deep learning architecture, shown in Figure 3.3, which is optimized to capture morphological and architectural features on varying scales with high efficiency and has been shown to achieve human-level prediction capability on the ImageNet dataset.[144] For the first and second stage models, we randomly assigned 30% of the 28 and 257 whole slide images, respectively, to held-out test sets used for model validation. Deep learning models and training routines were developed in Keras with Tensorflow backend[121] and trained undergoing cyclic learning



Network	k	l	m	n
Inception-v4	192	224	256	384
Inception-ResNet-v1	192	192	256	384
Inception-ResNet-v2	256	256	384	384

Figure 3.3: Learning architecture of the Inception v4 model illustrating the various sub-block layers employed in the overall design.

rates[145] with base learning rate of 0.001 and using the Adam optimizer.[46] To mitigate learned bias due to class imbalance, we utilize training data generators designed specifically to class-balance with over-sampling each batch of training. Models were trained from scratch on NVIDIA V100 GPUs made available through the Exacloud HPC resource at Oregon Health & Science University. The code used to generate the results and figures is publicly available.[2]

3.3 Results

3.3.1 Quantitative Localization of Liver Cancer in Whole Slide Images

The first-stage model is a tumor tile binary classifier that generates a prediction between 0 and 1 for each tile in the dataset in which a 1 corresponded to perfect confidence that

[2] `www.github.com/schaugf/NEMO`

a tile was of tumor tissue and in which a 0 corresponded to perfect confidence that the tile was of normal or stomal tissue. This model achieved an area under the precision-recall curve of 0.791 under the receiver operator characteristics curve which was sufficient to establish good correlation ($R^2 = 0.96$) between clinical estimation and our model's estimates of tumor purity (percent tumor in the whole slide image) as shown in Fig. 3.4. Further, visual comparisons between the pathological tumor annotation and our model's predictions illustrate spatial concordance between the drawn tumor-bounding mask and our model's predictions. Once trained, the tumor-region identifying model was deployed on the entire remaining dataset to include only tiles containing cancerous tissue. Several practical considerations contribute to our model's failure to perfectly reflect pathological annotation, including damaged tissue, necrosis, and stromal regions in the whole slide image. Because our approach in this case is limited to 28 WSIs, we anticipate greater data volume would improve robustness of our model to these and other tissue-specific morphological features.

3.3.2 Quantitative Whole Slide Image Classification of Metastatic Origin

After the first stage identifies regions of the H&E images that are tumor, the second stage model learns to classify those tiles according to their metastatic origin. A second Inception v4 deep neural network was designed to generate a three-class prediction for each tile in the training set as belonging to either a colonic adenocarcinoma, gastrointestinal stromal, or neuroendocrine carcinoma. Whole slide image predictions predictions aggregated across all corresponding tumor tiles achieved an F1 score of 0.875 on the held-out testing set of WSIs, having failed to correctly classify 5 out of the 51 held-out testing samples. Class-specific statistics shown in Table 3.1 quantify classification performance metrics for the

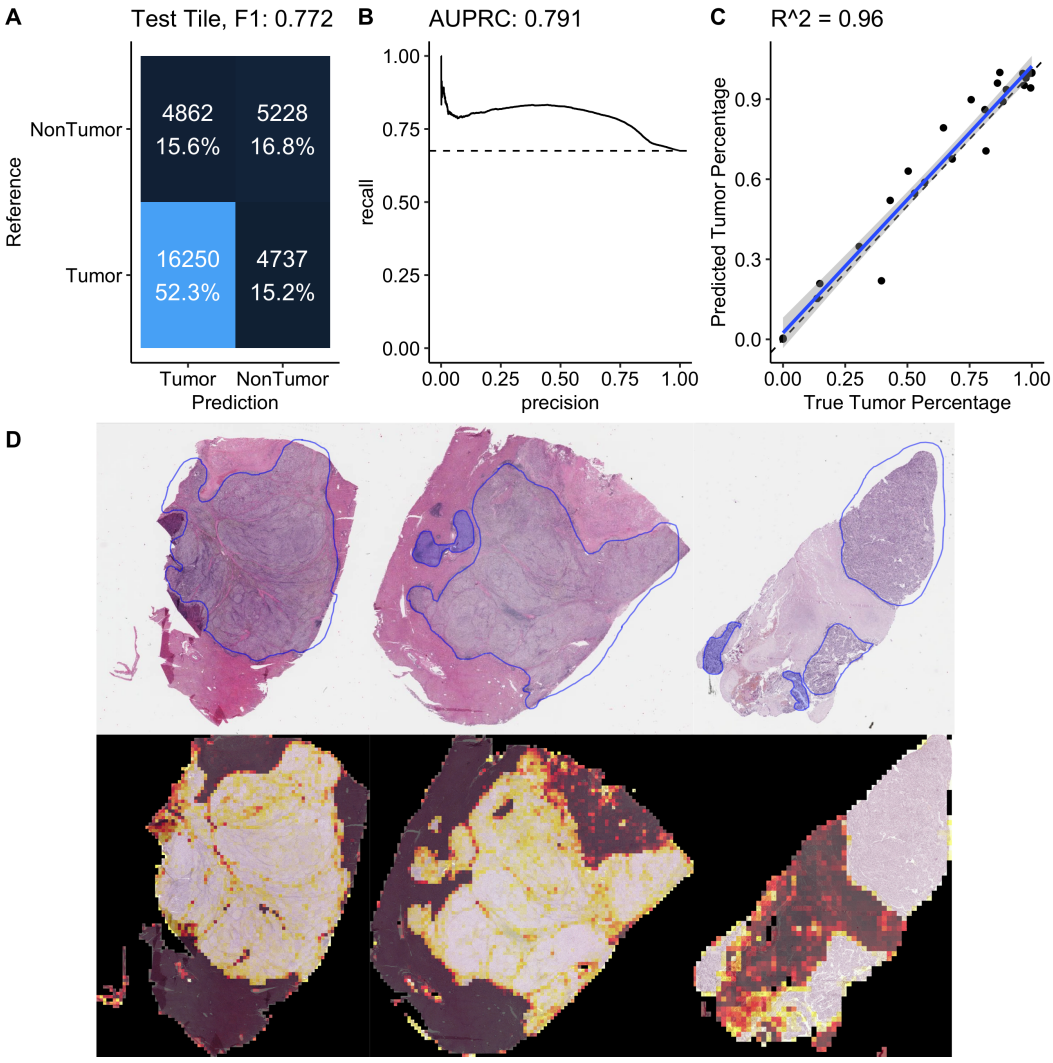


Figure 3.4: (A) Confusion matrix from the held-out testing set for a tumor/non-tumor predictive model illustrating F1 score of 0.772 in the classification task. (B) Precision-recall curve with area under the curve of 0.791. (C) Comparison between the true tumor purity in the sample inferred from the pathological annotation (x-axis) versus the inferred tumor purity from the model's output (y-axis) with strong correlation ($R^2 = 0.96$) (D) Three examples from the held-out testing set with pathological annotation of tumor regions outlined in blue (top) and corresponding model predictions estimating regions of whole slide images that contain tumor tissue (bottom) illustrating concordance between the pathological annotation of tumor region with the outcome of our model. In these illustrations, a brighter color intensity corresponds to higher probability that the underlying tile was labeled as being of tumor by the trained model.

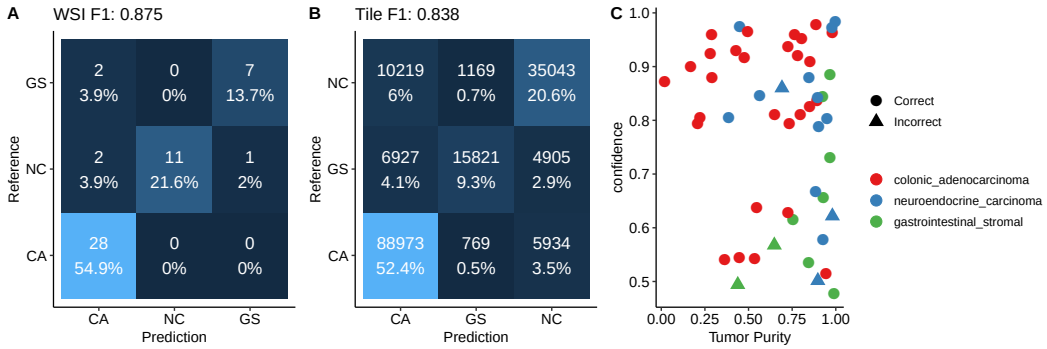


Figure 3.5: (A) Confusion matrix of WSI prediction on a held-out test set. (B) Confusion matrix of tile-based predictions. (C) failure cases with respect to the inferred tumor purity (percentage of the whole slide image that contains tumor tissue) in the sample on the x-axis (fraction of tiles predicted to be tumor) and the model's output confidence in its prediction on the y-axis.

metastatic origin prediction model. Confusion matrices of both WSI and per-tile predictions are shown in Figure 3.5. In this example, per-tile classification accuracy of 82.3% and per-WSI accuracy of 90.2% were achieved.

Several technical factors were associated with incorrect predictions, including slide blurring, tissue folding, and low tumor purity. Our model's confidence was lower for samples that it incorrectly classified, as shown in Fig. 3.5, though one sample was incorrectly classified with 86% confidence which was driven by misclassified stromal tissue present in the H&E slide. Individual tiles associated with highly confident predictions for each class are shown in Fig. 3.6. Pathological inspection of these tiles suggests that tiles associated with highly confident class predictions present pathological features that guide diagnoses, as the first row contains tiles presenting features associated with primarily spindle-type gastrointestinal stromal tumors and the third row presenting typical well-differentiated neuroendocrine carcinomas. The first two images in the second row represent dirty necrotic tissue which, among the three diseases under consideration, tends to be associated with colonic adeno-

Table 3.1: Class-specific statistics of both the tumor identification and three-way origin classification task

	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall	F1
Tumor Identification	0.77	0.52	0.77	0.53	0.77	0.77	0.77
Colonic Adenocarcinoma	1.00	0.83	0.88	1.00	0.88	1.00	0.93
Neuroendocrine Carcinoma	0.79	1.00	1.00	0.92	1.00	0.79	0.88
Gastrointestinal Stromal	0.78	0.98	0.87	0.95	0.88	0.78	0.82

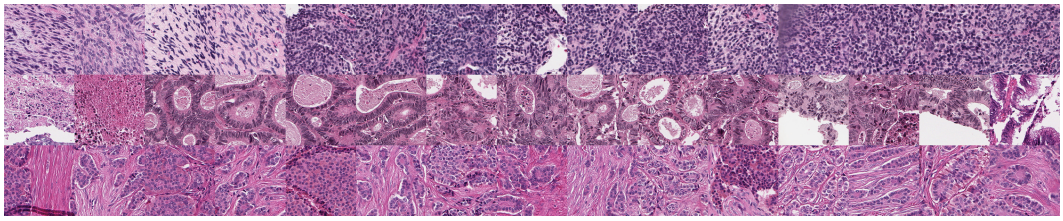


Figure 3.6: Example tiles correctly classified by the model with high confidence in which each row is a distinct class (gastrointestinal stromal, colonic adenocarcinoma, and neuroendocrine carcinoma in rows 1, 2, and 3, respectively).

carcinomas. However, this type of feature is not explicitly associated with cancer, and so should be interpreted with caution. Importantly, this approach obviates the need for pathological region annotation beyond what was required to train the first stage model.

3.3.3 Clinical Benchmark Comparison Study

A study was developed to benchmark our approach to clinical practitioners. This study recruited three board-certified pathologists to independently classify each of the 51 whole slide image samples in the held-out test set according to their metastatic origin. Each participant independently incorrectly classified 3, 4, and 5 samples each, while our neural network model missed 5 samples from the held-out test set. Table 3.2 summarizes the

Table 3.2: Slides mis-classified by either the model or at least one pathologist (GS: Gastrointestinal stromal; CA: Colonic adenocarcinoma; NC: neuroendocrine carcinoma). Instances of misclassification are highlighted in bold text.

Slide Alias	Ground Truth	Model	Path1	Path2	Path3
101	CA	CA	CA	CA	**NC**
102	CA	CA	CA	CA	**NC**
103	CA	CA	CA	**NC**	**GS**
104	CA	CA	**NC**	**NC**	CA
105	GS	**CA**	GS	GS	GS
106	GS	GS	GS	GS	**NC**
107	GS	**CA**	GS	GS	GS
108	NC	**GS**	**GS**	NC	**GS**
109	NC	**CA**	NC	NC	NC
110	NC	NC	NC	**CA**	NC
111	NC	**CA**	**CA**	**CA**	NC

eleven samples that were missed by either the model or by at least one pathologist and their respective predictions. Interestingly, only two of the mis-classified samples by the model were correctly classified by all three pathologists. Table 3.7 illustrates a selected sample classified correctly by the model and all three pathologists, a sample missed by the model that the pathologists all got correct, a sample missed by both the model and at least one pathologist, and a sample for which the model was correct but at least one pathologist made an incorrect classification. All examples illustrate the raw H&E image and three heat maps generated by the model for each of the three-way predictions in which a brighter color corresponds to a higher confidence in the model's prediction for each class. Importantly, predictions are only available for tiles that the first-stage of our model classified as tumor tissue, as non-tumor tiles were filtered out of the metastatic origin prediction task. Although the failure cases are diverse, probabilistic overlays of metastatic origin prediction may facilitate faster and more efficient examination of these tissue sections in clinical decision making processes.

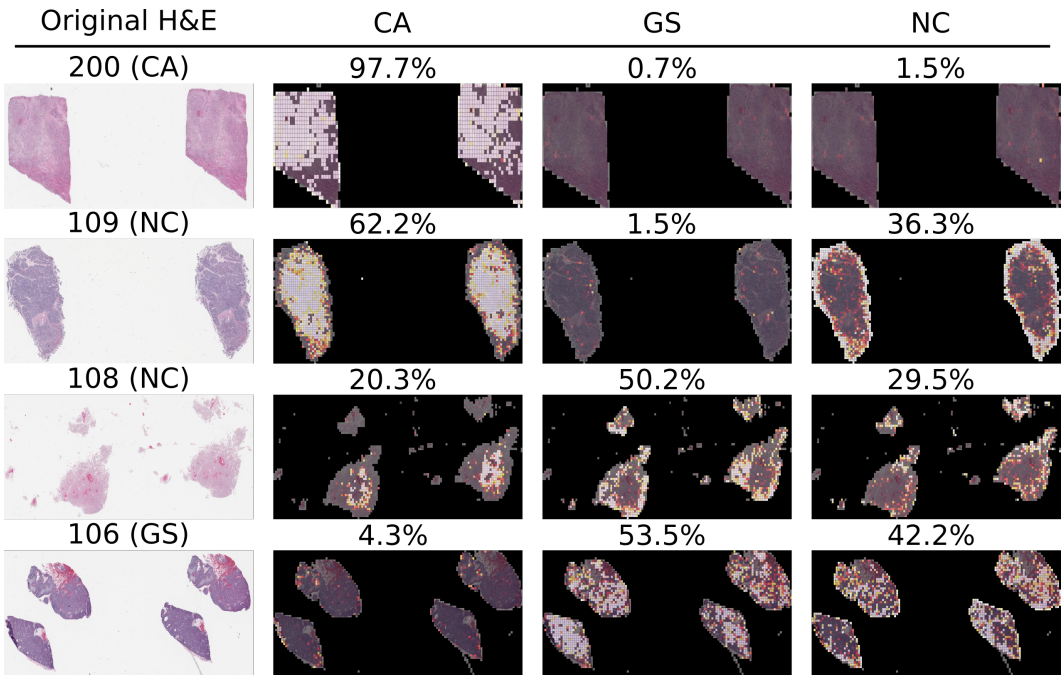


Figure 3.7: Example mis-classified H&E slides with associated annotations from the second stage model illustrating spatially-resolved localized predictions of metastatic origin. In these example images, brighter colors are associated with more confident class-specific predictions. First row: sample correctly predicted by the model and all three pathologists. Second row: sample missed by the model that all three pathologists got correct. Third row: Example missed by both the model and at least one pathologist. Fourth row: example missed by at least one pathologist that the model got correct. (GS: Gastrointestinal stromal; CA: Colonic Adenocarcinoma; NC: Neuroendocrine Carcinoma). The complete dataset of all high-resolution whole slide images and their associated colored prediction heat maps are available upon request.

3.4 Discussion

This work presents a deep learning based approach designed to infer the origin of metastatic liver cancer using a two-stage serial model composed of a first model trained to identify tumor from non-tumor within H&E sections of metastatic liver tissue based on pathologists annotation and a second stage model that learns to predict the metastatic origin of individual patches of tumor tissue and aggregates those results into predictions over WSIs. We illustrate through a clinical benchmark comparison that our approach is within perfor-

mance criteria of board-certified pathologists, suggesting that these types of systems may be capable of generating rapid, first-pass assessments of metastatic origin in the absence of detailed clinical information or comprehensive molecular profiling assay. We believe this type of data-driven visualization augmentation provides an additional layer of information that may facilitate the speed and ease of generating final decisions by clinical care providers.

Although these results illustrate feasibility of our approach, several significant limitations remain. Principally, this analysis was data-limited to only three most-prevalent sources of metastatic origin when in practice metastases can and do originate from a broad variety of biological sources. Secondly, we observe that the first stage model may be inflexible to alternative sites of metastatic tissue. Instead of training a model to identify tiles containing cancer tissue in liver, a more generalizable model may be trained on a broad diversity of primary cancers and regularized appropriately to identify cancer independently of the host tissue. Third, although our model was shown to perform similarly to board-certified pathologists, we have not thoroughly considered the manner by which these types of deep learning models might optimally improve current work-flows of practicing pathologists.

We believe that robust translation of deep learning systems such as the one presented in this paper may continue to supplement and augment clinical decision-making processes dependent on medical image analysis. The degree to which first and second stage models are generalizable would likely be improved with additional training data. Presently, this study was limited to a few hundred whole slide images in total for which pathological annotations were made available for only a few dozen. Although practical logistical issues prevent high-throughput annotation collection and processing, we believe that for this and similar types of systems to reach their full potential, robust integration of current bio-bank

and other data repositories with engineered data-processing pipelines must be established to facilitate rapid and reproducible research. While this approach is still in early stages, we nevertheless remain optimistic that future developments of computer vision systems may significantly contribute to improving the efficiency and efficacy of pathological interrogation of metastatic patient tissue. Future directions will seek to evaluate whether metastatic class boosting is achievable using primary tumor as analog training data, which is explored in greater depth in Chapter 5.

Disclosures

No conflicts of interest, financial or otherwise, are declared by the authors. This work was supported in part by the National Cancer Institute (U54CA209988), the OHSU Center for Spatial Systems Biomedicine, the Knight Diagnostic Laboratories, and a Biomedical Innovation Program Award from the Oregon Clinical & Translational Research Institute. We extend our thanks to the staff at the OHSU Knight BioLibrary for their support in data access and dissemination. Further, we gratefully acknowledge the resources of the Exacloud high performance computing environment developed jointly by OHSU and Intel and the technical support of the OHSU Advanced Computing Center.

Chapter 4

DeepHAT: Deep Histology Annotation Tool

Everything that can be invented,
has been invented.

Charles H. Duell
Commissioner,
U.S. Office of Patents (1899)

This work presented in this chapter has not been published and remains under active development. A tentative white paper and technical report has been circulated to core collaborators Hassan Ghani, Erik A. Burlingame, Guillaume Thibault, Christopher Corless, and Young Hwan Chang.

Abstract

The proposed method is predicated upon the observation that feature representations of clinically relevant tissue labels across various whole slide images is highly conserved. This work illustrates that unsupervised deep learning methods are capable of capturing morphological similarities in an interactive feature space that can be manually annotated. This work compares region annotations manually generated by an expert pathologist at whole-slide resolution with annotations generated against an interactive feature manifold. This work describes an open-source annotation tool that incorporates a semi-supervised learning approach that retains an expert pathologist in the loop. The tool is designed to learn meaningful feature space projections of tiles samples from whole slide histology and present the projections in a web-based application that enables a user to manually select, inspect, and annotate clusters of tiles that share similar morphological features. Validation experiments illustrate basic utility for composing annotations of a single-slide and multi-slide setting.

4.1 Introduction

Acquiring intra-slide annotations of whole slide histology is a major bottleneck in digital pathology, and yet because whole images are often very large, annotations are often necessary to refine class labels of heterogeneous regions within a single image. The size of whole slide images generally presents an intractable computer vision problem, so slides are commonly tiles into fixed-sized patches which are more easily computed on. Because morpho-spatial features of histopathology are generally conserved both within and across different whole slide images, the annotation process of identifying semantically-similar tissue features may be highly redundant; the features associated with hemorrhage or fat, for example, are generally conserved so as to enable a pathologist to engage in pattern-matching behavior to identify regions or features of interest within whole slide images. This work seeks to accelerate the annotation process by first grouping patches of whole slide histology across multiple samples based on morphological similarity and facilitating annotations of a composite representation. We seek to validate the approach by comparing annotations collected by this novel method to patch-level annotations generated by a trained pathologist on whole slide images.

Segmentation and annotation of biological images have been the subject of considerable attention in recent years, though the application of computer vision techniques to identify cell boundaries has been undertaken by various research groups for the better part of a century.[146] Several tools have been introduced to address high-throughput annotation of biological data. CellProfiler is a specialized tool designed for single-cell segmentation from diverse biological data types that incorproates deep learning methods.[147] Ilastik utilizes

semi-supervised annotation to iteratively update random forest models to predict regions of tissue that match annotations provided by the user in a near real-time manner.[148] Deep learning methods have emerged to leverage sparsely labeling of data to optimize categorization of imaging data. One-shot learning is one such example, in which information is borrowed from related classes to generate predictions of unseen classes by updating a probabilistic Bayesian model as new categories are observed.[149–151] Few-shot learning builds upon this idea by employing a common semi-labeled latent space to infer class identity of new data.[152,153] While few-shot learning presents a difficult computational problem, humans are in fact quite capable of one-shot classification.[154] Although model designs have shown to extend sparse labels to unseen data classes without the need for retraining from scratch, they rely on some trained data nevertheless. In cases of clinical histopathology, image labeling and annotation is expected to come from an expert pathologist, who through training develops reasonable models of prior expectations and distributions of a broad diversity of pathologies.

This work seeks to leverage the ability of a trained pathologist to inform labeled representations from scratch by annotating tiles samples from whole slide histology based on learned spatial features across a dataset. The approach described is intended to reduce the redundancy of annotating regions of interest that exhibit morphological similarities both inter- and intra-whole slide image, reduce overhead costs associated with loading and deloading whole slide images from a whole slide viewer, and be accessible to different users without the need for copying large datasets of histopathological tiles.

4.2 Data

4.2.1 Whole Slide Images

This work is supported by seventy eight hand-annotated whole slide images provided by the Knight BioLibrary shared resource at Oregon Health & Science University. For each image, a trained pathologist annotated tumor regions within the slide, such that image is compartmentalized into a binary class label as either tumor or non-tumor. Raw image data are provided as either `svs` or `scn` files from Leica or Apperio scanner hardware. Images are stored in raw format on the lustre storage system made available through the exacloud computing environment at OHSU.

The set of whole slide images provided by the BioLibrary is accompanied by a look-up table that relates each image to a set of clinical variables, including tissue of origin and the clinical diagnosis informed by the healthcare record. An expert clinical pathologist collated these variables into fourteen distinct target populations. These metadata are stored on the filesystem as flat comma separated value tables.

4.2.2 Whole Slide Manual Annotations

An expert pathologist was tasked with manually labeling tumor regions within whole slide images using the QuPath[1] software tool for quantitative pathology as shown in Figure 4.1.[61] The expert pathologist annotated 78 whole slide images sampled across the fourteen designated class labels. Custom tooling was developed in python and Apache Groovy to extract manually-annotated regions and store binary mask arrays for each whole slide image.

[1] `https://qupath.github.io/`

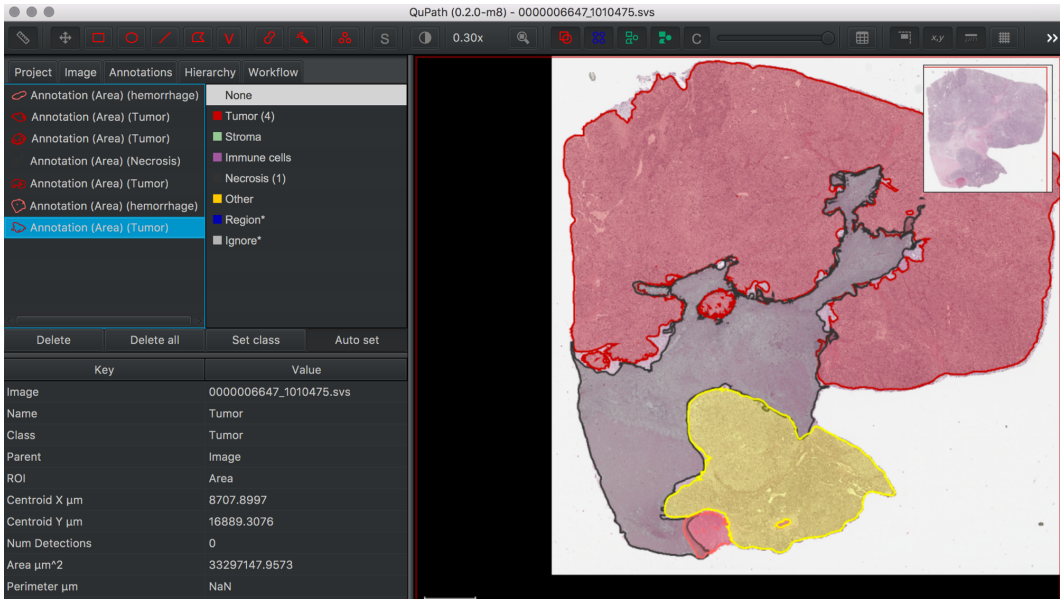


Figure 4.1: QuPath user interface of whole slide image undergoing expert annotation. In this example, unannotated regions are unlabeled and so are excluded from this analysis. Tiles that are ambiguously labeled are labeled according to the majority label in the tile.

Preprocessing of whole slide images includes linkers to generated annotation masks and enables patch-level labeling to indicate whether a sampled patch from a whole slide image is sampled from within or external to tumor tissue.[125]

4.3 DeepHAT Design

4.3.1 Whole Slide Image Preprocessing

Whole slide images are loaded into python through the OpenSlide[2] python API and preprocessed to first tile the whole slide such that each tile captures a uniform surface area across multiple slides and resolutions such that each tile captures 100 μm^2 and resized to 128×128 pixels. Background tiles are filtered out by removing tiles whose mean intensity is greater than 230 and less than 20. Remaining tiles containing tissue are color normalized by an

[2] `https://openslide.org/`

established method. (Macenko, *et al*). Each of N tiles from an image are concatenated and stored as numpy arrays if size ($N \times 128 \times 128 \times 3$) at an unsigned 8-bit color resolution.

Once each image is processed, a second routine loops through each array, loads them into memory, and stores them in a binary file format onto the disk in an HDF5[3] data storage format that allows for dynamic indexing directly from storage.

4.3.2 Histological Representation Learning

Similar tiles of histology should be clustered together to enable more rapid annotation of whole slide features across multiple images, which necessitates a similarity metric or quantitative representation of tiles on which to compute. A variational autoencoder (VAE)[49] was trained to learn meaningful representations of whole slide histology by encoding the learned histological features into an abstract feature space such that similar types share similar encoding representations. Model were designed with the PyTorch[4] library for deep learning and trained on a multi-GPU instance equipped with four NVIDIA V100 processors and 64 GB random access memory. Preliminary results suggested compelling learning characteristics given a batch size of 64 and a latent representation dimension of 32 floating-point values. A dataset composed of approximately 30k tiles were employed to train a model for 10 epochs with a base learning rate of 1e$-$3 updated by the Adam optimizer and with a learning rate scheduler designed to reduce the learning rate upon reaching a plateau by a factor of 0.1. Once training is complete, model weights and parameters are stored and each tile is independently encoded into a latent representation and saved as an $N \times D$ array containing N tiles across D learned latent variables.

[3] `https://www.hdfgroup.org/solutions/hdf5`
[4] `https://pytorch.org/`

4.3.3 Feature Manifold Projection

Feature encodings are visualized with either the tSNE[122] or UMAP[155] projection algorithms, both of which are designed to project the learned feature space into a two-dimensional space. In both cases, implementations are used from the RAPIDS[5] library which is built upon the CUDA[6] toolkit that enables utilization of GPU-computing architecture in performing operations. This approach projects every tile as a point in a two-dimensional space that is easily visualized in a scatter plot as shown in Figure 4.2 in which projected tile features are colored by the clinical whole slide image diagnosis. Morphological features associated with regions in this space are visualized with tile plots, which first conform the continuously-valued projection space onto a regular grid manifold onto which tiles are directly visualized, enabling confirmation by inspection that the VAE model learns meaningful representation of the tile dataset as shown in Figure 4.3. Although the unsupervised autoencoding approach does not perfectly segregate populations of tiles based on their clinical diagnosis, certain morphological features of the tiles tend to be clearly clustered in an entirely unsupervised manner. If structural features of interest were to be evident in this type of representation, then a pathologist may achieve significant efficiency when annotating this representation rather than each whole slide image independently. For example, this approach may not be sufficient for achieving high accuracy when annotating different cancer subtypes, but may provide an efficient tool for segregating tumor from non-tumor tissue or for removing undesired tiles that capture edge effects or technical artifacts.

[5] `https://rapids.ai/`
[6] `https://www.geforce.com/hardware/technology/cuda`

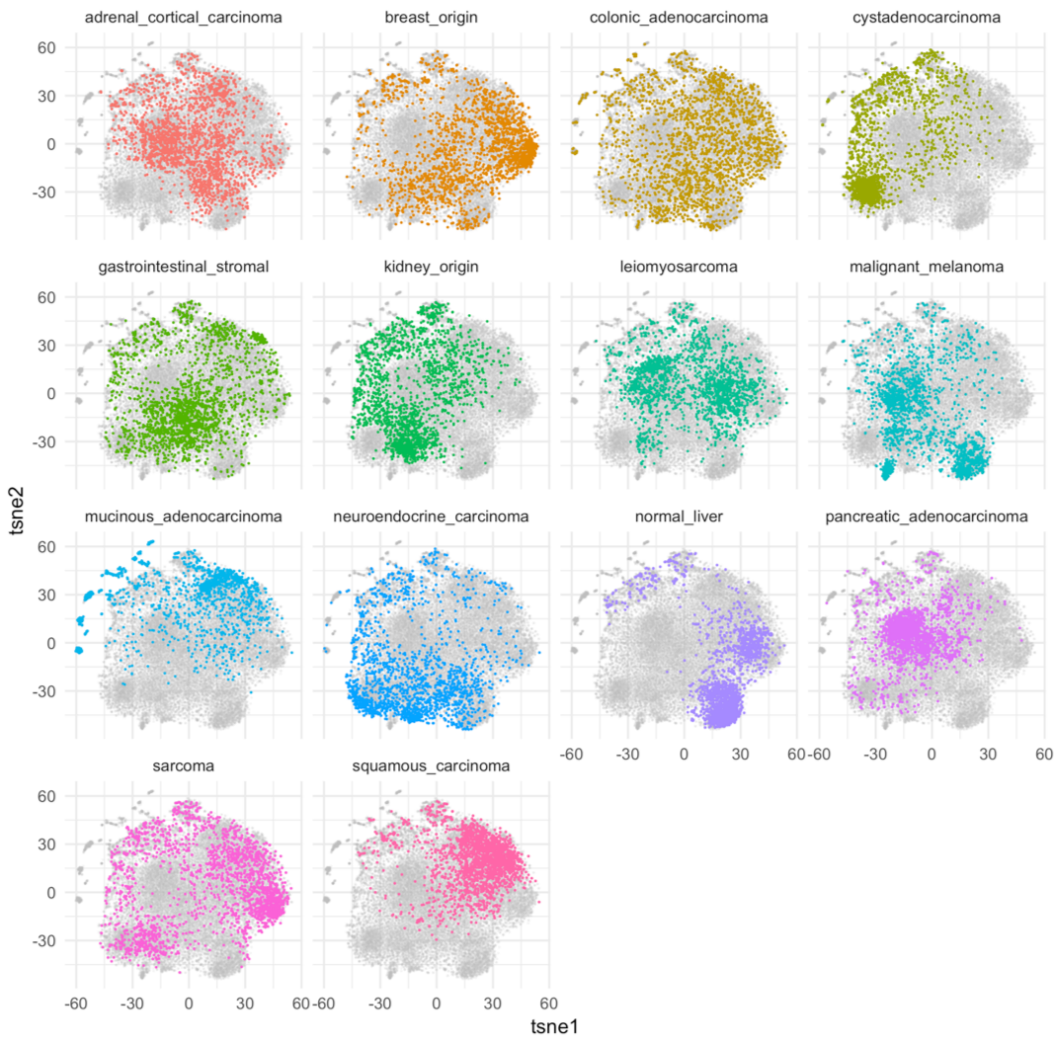


Figure 4.2: Unsupervised feature manifold projection of latent feature space representation of tiles samples from whole slide histologies of 14 tissue types. In this example, similar whole slide diagnoses tend to co-cluster, suggesting that unsupervised spatial features specific to metastatic indications are conserved across whole slide images. Normal liver features tend to tightly co-cluster which is expected, as normal liver histology has a regular and distinct spatial structure. Conversely, adrenal cortical carcinomas appear to be more dispersed which suggests greater variability in the learned spatial features associated with this metastatic cancer sub-type.

4.3.4 Interactive Annotation Application

An interactive interface to the projected feature space is necessary to facilitate the generation and storage of per-tile annotations of the learned feature space. The bokeh[7] library

[7] `https://bokeh.org/`

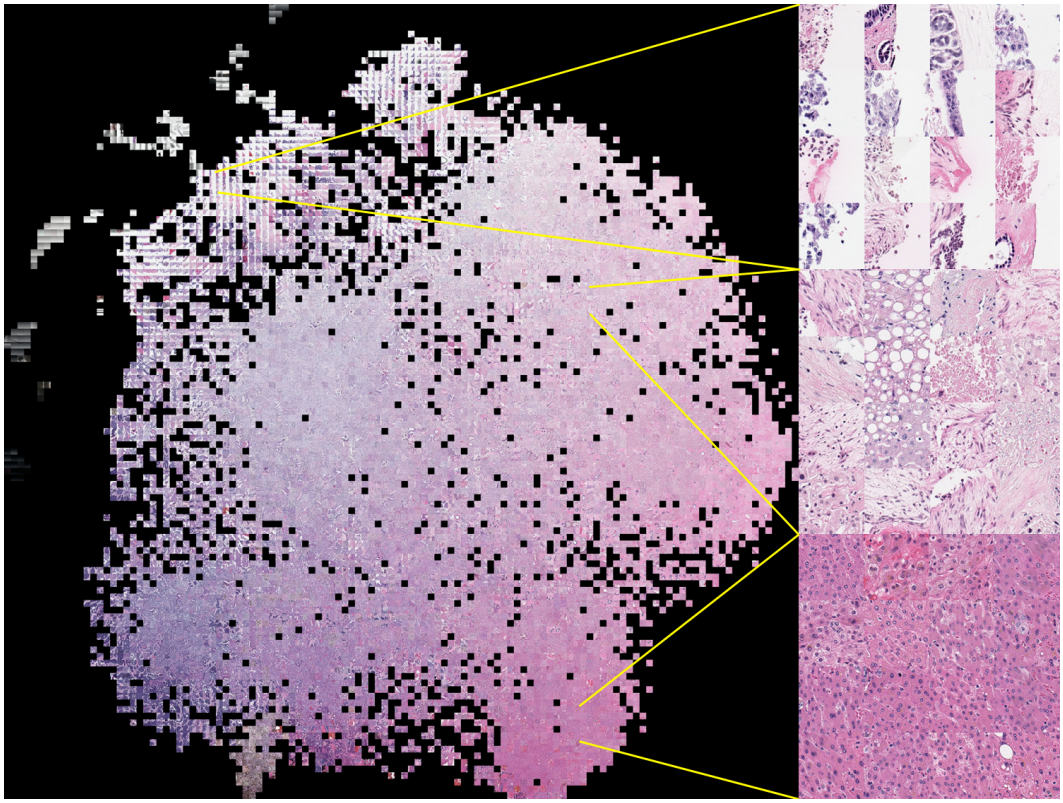


Figure 4.3: The point-projection plot shown in Figure 4.2 is here rendered by positioning tiles onto a coarse grid of the coordinate space. Callouts illustrate varying degrees of preserved morphological similarity. The top call-out illustrates a distinct region of feature space capturing edge effects largely independent of clinical subtype. The middlel callout illustrates a region of a high degree of morphological variability that includes tiles containing fat cells, stroma, and striated tissue. The bottom callout illustrates a region of high morphological homogeneity composed entirely of normal liver samples.

is designed to support server-based interactive data visualization and interaction and is utilized to incorporate the learned feature space projection with an engine to select and annotate individual data points. Importantly, because selections are made with small polygons in specific regions of the projection, a user is able to capture many tiles that share similar encodings and morphological features. The interactive scatter point plot incorporates user tools enabling a user to pan and zoom in or out. Once a selection of points is made, the user can query the tile dataset stored on disk to sample sets of tiles at full resolution to guide

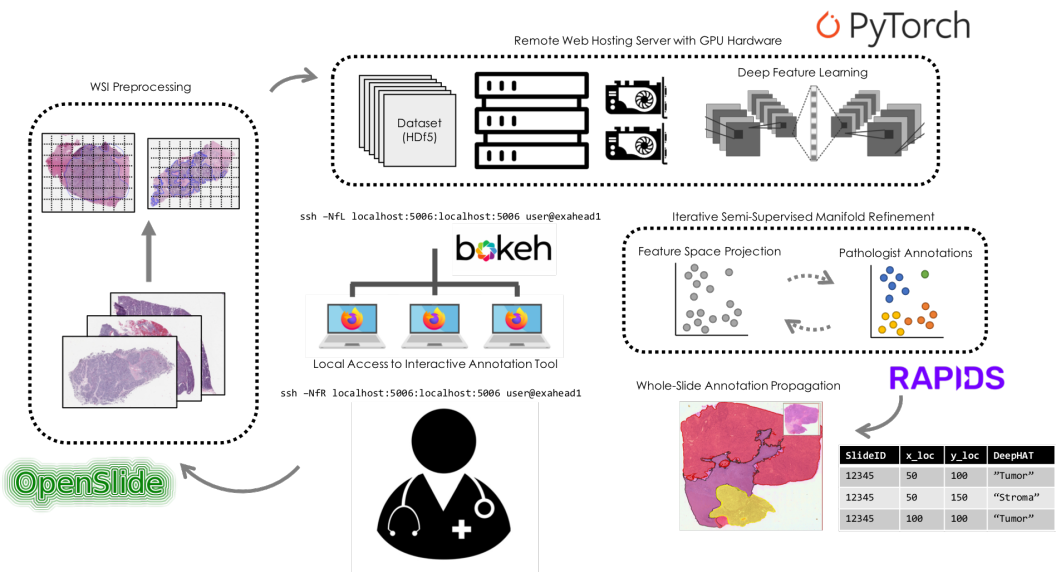


Figure 4.4: Annotation application computational overview

the annotation decision. The tool is equipped with a few basic selection options similar to those available in the QuPath tool, such as Tumor, Stroma, Hemmorhage, etc. However, the user may create their own annotation as a free text string, which the tool remembers and automatically appends to the drop-down list of pre-build annotations. For a given selection, the user can define a character string annotation which is automatically saved into a separate datafile that associates each tile to its annotation. The general processing work-flow used in the annotation is illustrated in Figure 5.1.

4.3.5 DeepHAT Accessibility

The tool is hosted on an instance of the exalogin partition on the exacloud computing system and configured to broadcast its interface on a specific port that remote users can forward to their local machine without necessitating download of large raw datasets. Remote users can access the visualizer by forwarding the bokeh server broadcast port to their

own computer, and can be forwarded with: `ssh -vNfL localhost:5006:localhost:5006 user@exalogin` Where `user` is the user name of an approved user on the Exacloud system. If a user is using a machine running the Windows operating system and does not have access to the secure shell command line, a freely available extension for the Chrome Web Browser, the Secure Shell App[8] has been tested and also works. Once the ssh tunnel is dug, the user can interact with the tool from their local web browser by navigating to `localhost:5006`. This approach thereby enables an expert pathologist to rapidly annotate hundreds or thousands of whole slide images stored on a secure remote server from their own local web browser.

The application consists of a few distinct components to guide the annotation process as shown in Figure 4.5. The left panel is highly interactive, allowing a user to zoom in and pan around the feature space embedding. A number of utilities are included that enable a user to select regions of the feature map, such as the small selection shown with blue squares near the center of the feature map. The right panel provides a zoomed-in view of selected tiles. In this example, the selected regions appears to contain primarily necrotic tissue as evidenced by the morphological features of the shown tiles. The bottom right panel allows a user to either select from a set of pre-defined annotated classes or enter in a new string to define the selection. Annotations are saved as a new column in a flat csv file, which is easily saved and reloaded into the DeepHAT tool for down-stream analysis or further review. Because the save format is simple, a user can easily share their annotations with other users who wish to modify the annotations generated. The inherent limitations of the number of data points that HTML can render in a single instance limits the number of

[8] `https://hterm.org/`

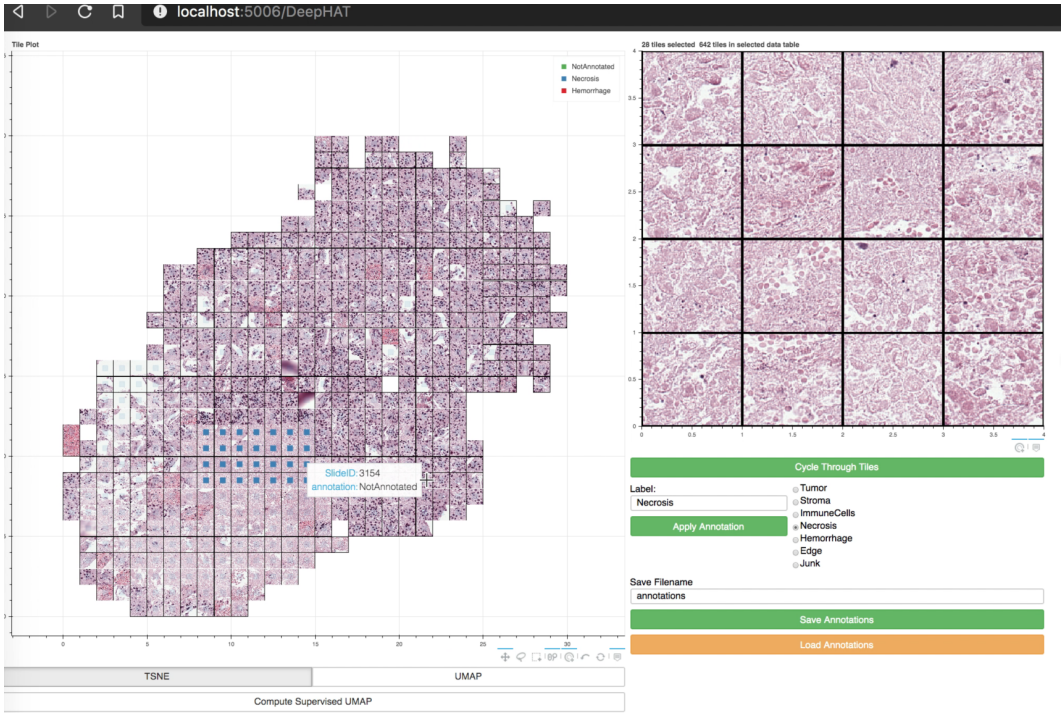


Figure 4.5: DeepHAT user interface to the web-based annotation application. The left panel is highly interactive and allows the user to zoom, pan, and select sets of tiles. The top right panel allows a user to cycle through each of the selected tiles. The bottom right utilities allow a user to select pre-defined annotation classes, enter their own, and save the resulting annotations to disc.

data points a user can annotate at once. Qualitative experiments suggest that the number of points rendered should be kept below ten thousand to retain fluid interactions with the annotation tools. Because the number of data points rendered in the annotator is limited, a mechanism is required to effectively propagate annotations to the remainder of the dataset.

4.4 Results

4.4.1 Single Slide Annotation

For illustration purposes, a single annotated histopathological whole slide image was first evaluated with DeepHAT to validate whether feature separability was observed and to

confirm that the annotation utility properly traced relevant tile meta-data for pseudo-whole slide image annotation. The shown sample is a whole slide image of adrenal cortical carcinoma from the NEMO dataset shown in Figure 4.6(A). The slide was tiled, normalized, encoded, and projected into 2-dimensions as shown in Figure 4.6(B) in which tiles are colored according to the annotations generated by a pathologist and demonstrate good separability between tumor and non-tumor tissue. Annotations of the feature manifold from the single image were generated by a trained expert using the DeepHAT tool, which were later post-transformed back into the original image, as shown in Figure 4.6(C), in which good concordance is observed between whole slide annotation and annotations made in DeepHAT. The confusion matrix of mis-classified tiles is shown in Figure 4.6(D), in which across 6384 tiles sampled from the image, the DeepHAT annotations aligned to their whole slide ground truth with 85% accuracy, 90% specificity, and 82% sensitivity. Zoomed-in callouts of the single whole slide image feature space projection illustrate local feature similarity as shown in 4.7.

4.4.2 Tumor Annotation from Whole Slide Primary Tissue

A second validation experiment sought to evaluate the reliability of whole slide annotations made with DeepHAT when annotating tumor tissue from normal tissue in primary cancers. A set of 78 whole slide images with hand-annotations generated in QuPath were tiled, labeled, and encoded using the VAE learning architecture described above. Figure 4.8 illustrates the learned feature space projection of the dataset illustrating the raw tiles atop tSNE embedding coordinates as well as point projections colored according to annotations generated in DeepHAT by an expert pathologist and colored by the annotations generated

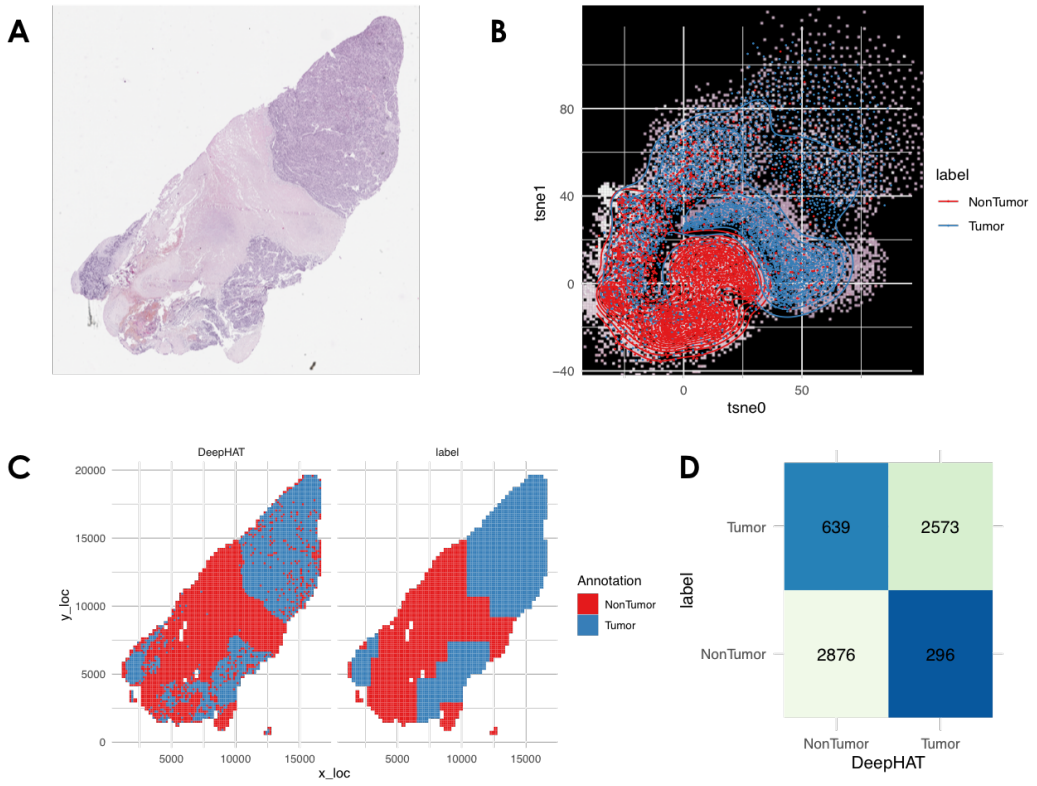


Figure 4.6: (A) Thumbnail of a single whole slide image. (B) Corresponding feature manifold projection colored according to the tile label provided by a pathologist at whole slide resolution. (C) Pseudo-whole slide image annotation is achieved by back-transforming the feature manifold into the original whole slide image coordinate system. (D) Confusion matrix of binary intersection of annotations generated at whole slide resolution and using the DeepHAT tool.

at whole slide resolution. The confusion matrix between tumor and non-tumor tiles labeled by DeepHAT and using QuPath illustrates an accuracy of 74.8%. The acceptable quality of annotation performance clearly depends on the questions being asked, but importantly, this result is achievable with merely 14 unique polygon annotations in the DeepHAT tool, amounting to about 10 minutes of a pathologist's time. This is in start contrast to the over 500 unique polygons generated across several dozen whole slide images that took several weeks of intermittent effort to generate. We also illustrate that per-tile errors may be restricted to certain whole slide images, suggesting that some images tissues may be more

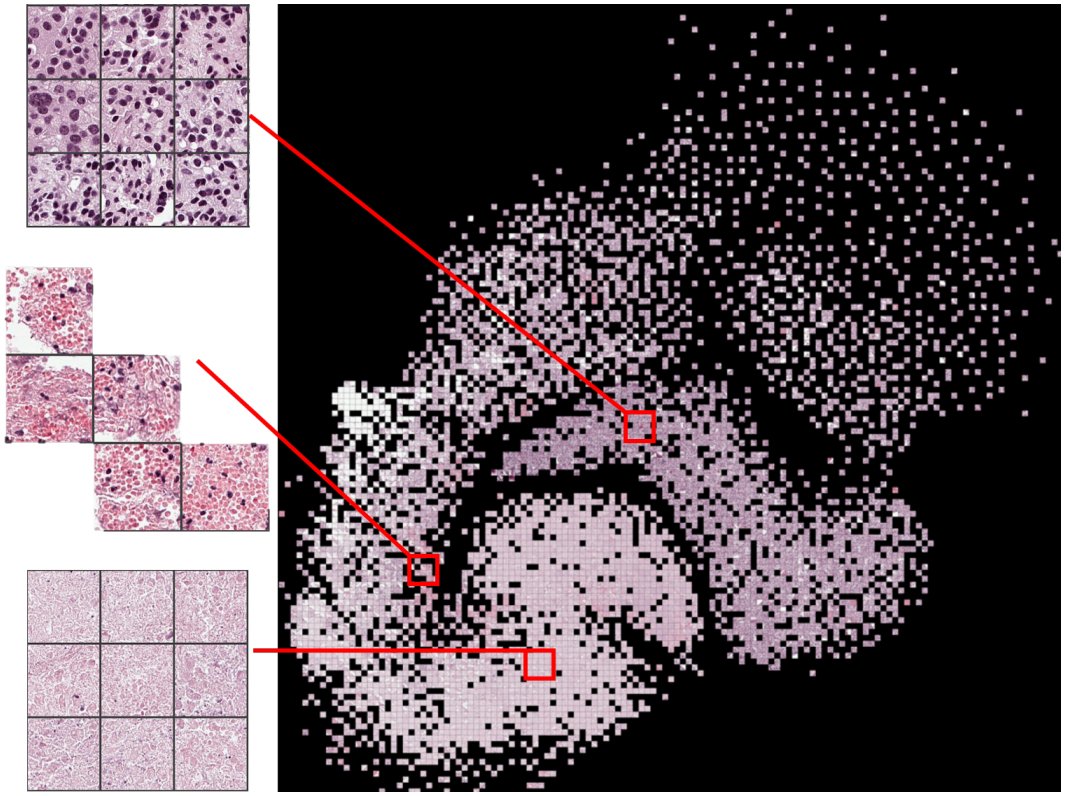


Figure 4.7: Histopathological feature space projection of a single slide. Callouts illustrate local feature similarity throughout the feature manifold. To generate this type of figure, the learned VAE feature representations for each of the tiles in the dataset are projected into two dimensions with the UMAP algorithm. Points are then fit to a grid of fixed dimension, and original tile images pasted over their coordinates.

or less difficult for an unsupervised learning approach to parse given morphological features in the samples tiles.

4.5 Discussion

This work presents an approach to accelerate whole slide annotation by learning feature representation of tiles of whole slide histopathology, embedding those features into two dimensions, and incorporating the resultant projection into a tool designed to acquire and store per-tile annotations. Although this approach holds promise for parallelizing image an-

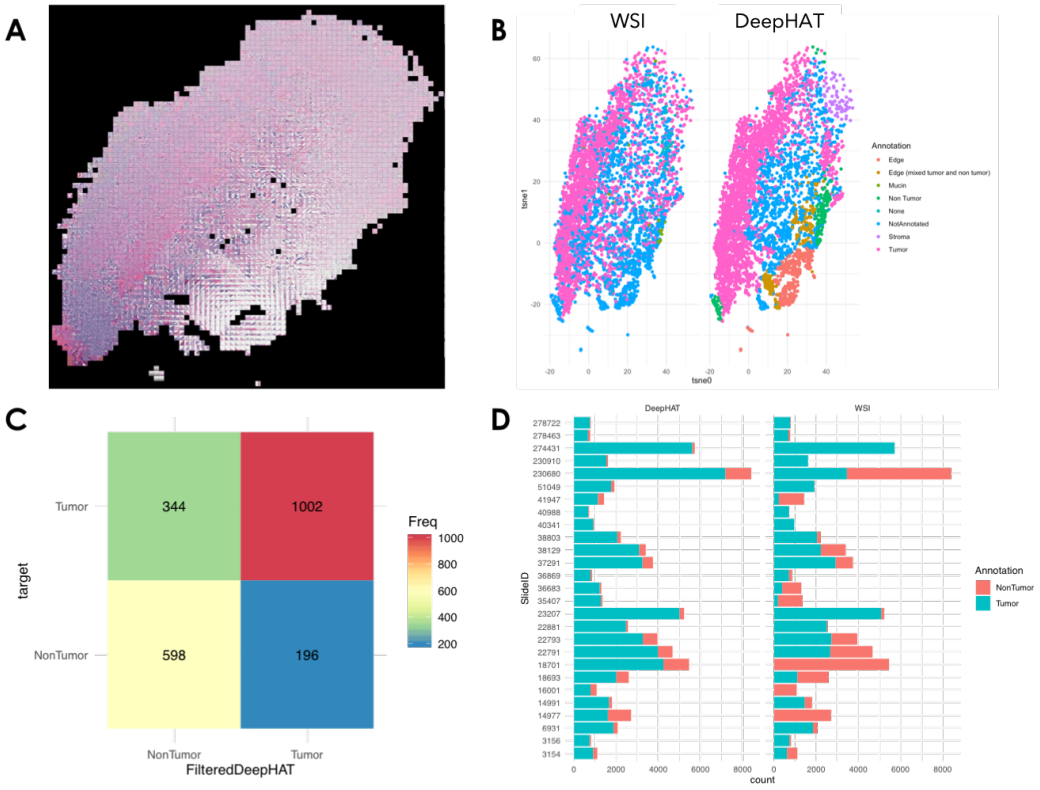


Figure 4.8: (A) Learned feature space projection of tiles sampled from 78 whole slide images of primary cancers. (B) Annotated feature manifold from the DeepHAT tool (left) and at WSI resolution (right) illustrates agreeable concordance between tumor and non-tumor labels. (C) Confusion matrix (in thousands of tiles) illustrating a per-tile accuracy of over 74% using 14 unique selected polygons in DeepHAT with respect to over 500 unique polygons across 78 whole slide images in QuPath. Annotation classes defined in (A) by the user are collapsed into tumor/non-tumor classes for purposes of binary comparison. (D) distribution of labels across several whole slide images illustrates varying degrees of concordance at whole slide resolution, suggesting that features from some whole slide images may be more challenging to meaningfully separate into tumor and non-tumor components.

notation, a number of limitations remain under-explored. Importantly, due to the rendering capabilities of the chosen library in HTML, a limited number of tiles are able to be rendered in a dynamic and interactive environment. Because the tSNE and UMAP algorithms project data into floating point numeric space, resultant projections are forced onto a grid to improve the ease of visualization. This results in a pile-up of tiles that are rendered in HTML, which is generally limited to fewer than 5000 in a single instance. Although

we explicitly designed a tile-viewer utility to allow a user to cycle through a selection, the purity of the stacks of overlapping tiles is variable within some regions of the feature map is variable. Whether this overlapping label variability is due to the coarse grid space or due to insufficient separation of tumor from non-tumor histologies likely depends on the type of data undergoing embedding, so inferring consistency of learned feature space embedding is an important direction of future research.

The most significant challenge to this approach is the operating assumption that unsupervised learning architecture is capable of learning meaningful representation of the processed tiles so as to enable quality annotations given the feature space projection. Although evidence suggests that these models may at times achieve compelling embeddings, issues regarding consistency remain, such that two tiles that look by eye to be highly similar are nevertheless distributed throughout different parts of the feature map. One potential solution would incorporate on-line semi-supervised learning so as to leverage whatever annotations have been provided to iterative refine the feature space projection to best delineate annotated classes. Although the tool has a few of these features already included, such as a semi-supervised UMAP algorithm and a linear discriminant analysis transform feature, their performance still relies on the feature space representation to stratify classes of interest. For example, if two tiles are each labeled tumor and non-tumor, but are embedded into identical regions of feature space, then no degree of semi-supervised projection will be able to distinguish the two.

Finally, although the tool is currently operational on a closed internal GPU-equipped server, the system still is slow in terms of time required to train an auto-encoder model and render the tile projection once loaded. This suggests the possibility of employing

previously trained auto-encoder networks to do the feature extraction without needing to retrain a model. Although this type of transfer learning has shown promise in many other domains, it is limited to the inherent biases of the datasets used to train the preliminary model. Future efforts will seek to incorporate more robust software engineering principles to resolve these and related bottlenecks in pre-processing, rendering, and hosting. Taken together, the DeepHAT approach represents an unsupervised method capable of achieving significant improvements in efficiency of a pathologist tasked with annotating large numbers of whole slide images.

Chapter 5

Primary-Metastatic Transfer Learning

> Never estimate what may be accurately computed... Never guess what may be estimated... Never guess blindly...
>
> *Julian Bigelow*
> *Chief Engineer of the IAS Machine,*
> *The world's first digital computer*

This work presented in this chapter has not yet been published, but is done with core collaborators Hassan Ghani, Erik A. Burlingame, Guillaume Thibault, Christopher Corless, and Young Hwan Chang titled "Histological transfer learning from primary to metastatic whole slide images".

Abstract

Accurate diagnosis of metastatic cancer is essential for identifying optimal treatment control strategies to halt further spread of metastasizing disease. While pathological inspection aided by immunohistochemistry staining provide a valuable gold standard for clinical diagnostics, deep learning methods have emerged as powerful tools for identifying clinically relevant features of whole slide histology relevant to identifying a tumor's metastatic origin. This works seeks to evaluate a deep learning system trained to classify secondary tumor histology based on presented spatial histopathological features of both primary tumor and their metastatic counterparts by transferring a learned model trained on primary cancer into the metastatic setting. A fourteen-way classification model designed to infer the metastatic origin directly from whole slide images of H&E stained tissue achieved mean class-specific areas under receiver operator characteristics of 0.779, which outperformed comparable models trained on only images of primary tumor (mean AUROC of 0.691) or trained on only images of metastatic tumor (mean AUROC of 0.675). We further sought to evaluate whether similarity of morphological features from primary and metastatic tumors that share a common cancer type. In this work, we identified a correlation between the mean Kullback-Leibler divergence between primary and metastatic samples from each class and the model's class-specific AUROC of 0.317, suggesting only a slight association between inferred dissimilarity between primary and metastatic tumors of similar type and the degree to which a model was successful in classifying them according to their class in a transfer learning setting.

5.1 Introduction

We previously illustrated a learning system trained to classify whole slide images of metastatic liver cancer according to the cancer's metastatic origin,[125] which is discussed in detail in Chapter 3. That work was limited to three of the most commonly recurring classes of metastatic cancer due in part to limited data availability of other classes of metastatic cancers, while this work seeks to leverage morphological spatial properties of primary tissue to enhance the performance of a classification model trained to infer the origin of secondary metastatic cancer based on histopathological presentation in digital whole slide images. Although similarities between primary and metastatic tumors have shown striking similarity in gene expression,[97] growth characteristics,[156] and chromosomal rearrangement,[157] presently no quantitative measurement of primary-metastatic similarity has been extended to the morphological presentation of cancer within histopathological images. Our proposed hypothesis is that if primary and metastatic cancers exhibit morphological similarities in whole slide images, then incorporating primary cancers into training a computer vision system to classify metastatic cancers may confer a modeling advantage.

Training a model in one setting and transferring its learned parameterization into a different setting is an example of transfer learning. These approaches have demontrated robust capacity for boosting model performance,[158] and have found wide use in the field of deep learning for computer vision applications.[159,160] This work extends previous analyses by evaluating the degree to which a computer vision model generalizes to unseen samples of whole slide metastatic cancer by training on only primary cancers, only metastatic cancers, and by first training on primary and transferring the learning model to retrain on metastatic

samples. Further, this work evaluates the divergence between primary and metastatic cancers of different types within learned unsupervised morphological feature space, and draws connections between the degree to which primary and metastatic cancers are dissimilar and how well models generalize to correctly predicting metastatic origin of different cancer types.

5.2 Methods

5.2.1 Computational Pipeline

An overview of the computational steps taken in this work is presented in Figure 5.1, which broadly separates the study objectives into two components. The first portion of the study seeks to pre-train a neural network classifier on whole slide images of primary cancers, while the second portion seeks to transfer the learned classification model into the metastatic setting. Both steps involve similar processing pipelines but with complementary objectives. Because whole slide images are large and heterogeneous, they may contain both tumor and non-tumor tissue. This mixture of tissue types has been shown to confound the degree to which classification models are generalizable, and so both components of this study require a model to identify tumor tissue from non-tumor tissue within the whole slide image. Both models are informed by manual annotation of whole slide images done by a board-certified pathologist. Annotated whole slide images are divided into training and test sets, and independent binary classifiers were trained to classify tumor and non-tumor tissue. Once validated, these preliminary filter models are deployed onto their respective whole datasets to filter out normal stromal tissue from the tiled datasets such that the

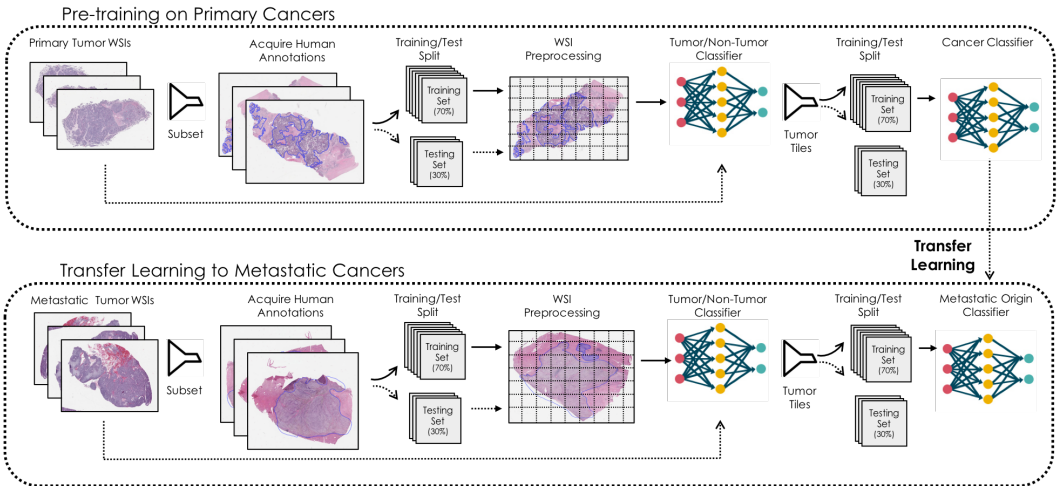


Figure 5.1: Transfer learning concept designed to leverage morphological and spatial features of primary cancer to infer the metastatic origin of secondary cancers.

resultant datasets are composed primarily of cancer tissue. Secondary models are then trained to correctly classify the tumor tiles according to their tissue of origin as informed by the clinical diagnostic record. Samples were chosen to reflect 14 common sites of metastatic origin to include both primary cancers and associated metastatic cancers of the liver. These whole slide images are similarly divided into training and testing sets and utilized to inform classification models. This work sought to evaluate the role of a transfer learning paradigm to evaluate whether a model trained just on primary cancer can infer a metastatic cancer's origin and whether a model trained on primary cancer but transfered into the metastatic setting for additional re-training outperforms the baseline approach.

5.2.2 Whole Slide Pre-Processing

This work employs dataset composed of 324 whole slide images of metastatic cancers and 344 whole slide images of primary cancer collected from the Knight BioLibrary and Knight Diagnostic Laboratories at OHSU. Each whole slide image read and processed through the

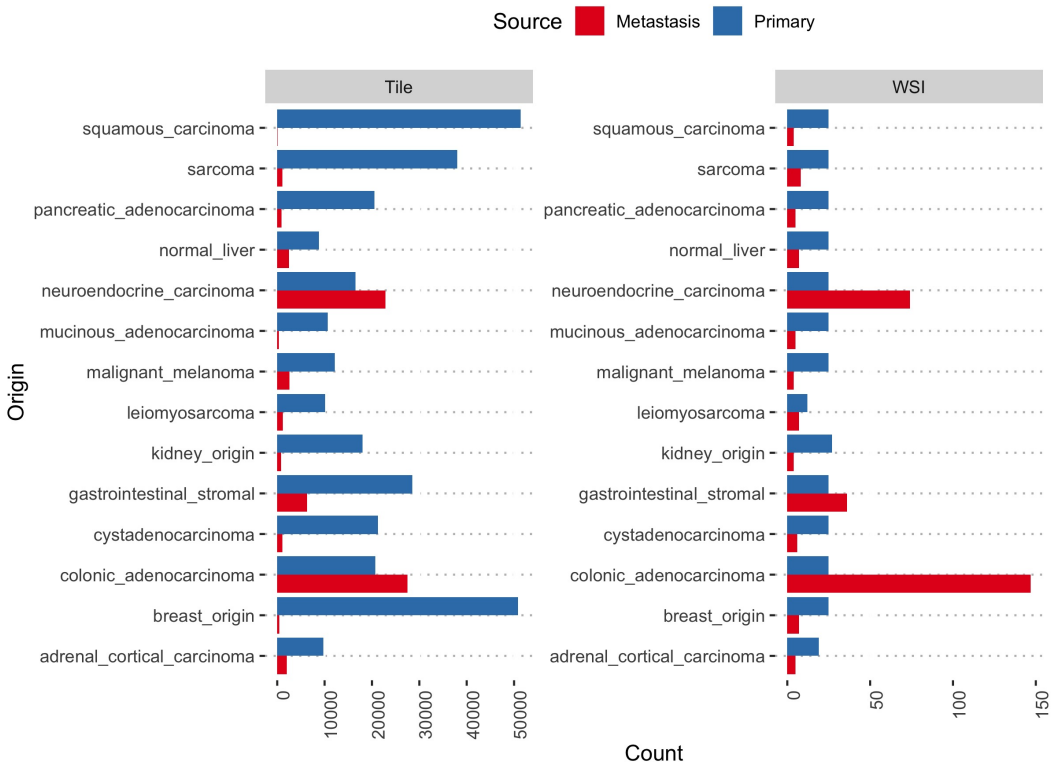


Figure 5.2: Tile and WSI count across the fourteen annotated classes from both primary and metastatic datasets illustrating a more uniform and representative training set across the fourteen classes of interest

OpenSlide python API which is used to divide the large images into non-overlapping tiles $128 \times 128 \times 3$ pixels wide that cover 100μm square. Tiles containing predominantly white background light were filtered out and the remaining tiles were normalized.[86] Annotation tables provided by the Knight BioLibrary associate each whole slide image with its tissue of origin informed by the clinical diagnostic record. The total counts of both whole slide images and individual tiles collected for both the primary and metastatic dataset are shown in Figure 5.2. The inherent class imbalance in the metastatic whole slide training set presented a limiting factor for previous work, which is intended to be overcome in part by a transfer learning paradigm.

5.2.3 Learning System Architecture

This study utilizes the ResNet50 learning architecture, shown in Figure 5.3, which is a powerful learning architecture that has been widely applied to challenges in digital pathology.[79,130,132] We incorporate the baseline ResNet50 model and modify its output layer to contain 14 nodes, each corresponding to one of the fourteen sites of tissue origin, and a softmax activation function such that the vector of output values sums to one and thereby enabling a probabilistic interpretation of the model's output. We trained a single ResNet50 model for 10 epochs with a learning rate of 0.001 and batch size of 32 on the complete training set of whole slide images from either primary or metastatic cancers. In all cases, the Adam optimizer was employed with a learning rate scheduler designed to decimate the learning rate at the end of each epoch, and class-balancing data loaders are used to maximize class diversity with each batch of training data. To evaluate the transfer learning approach, a third model is trained for 5 epochs on primary cancer and 5 epochs on metastatic cancer.

5.3 Results

5.3.1 Prior First-Stage Tumor Identification

Previous work introduced an approach for training a first-stage model to filter our normal tissue of metastatic liver examples by learning to classify tiles given expert pathologists' annotation of tumor and non-tumor regions of the images to remove normal tissue and other stromal regions from confounding the model's prediction. Similarly, a two-stage model is proposed for the primary cancer setting in which a binary classification model is trained to remove normal tissue. First, we evaluate the performance of the first stage model trained

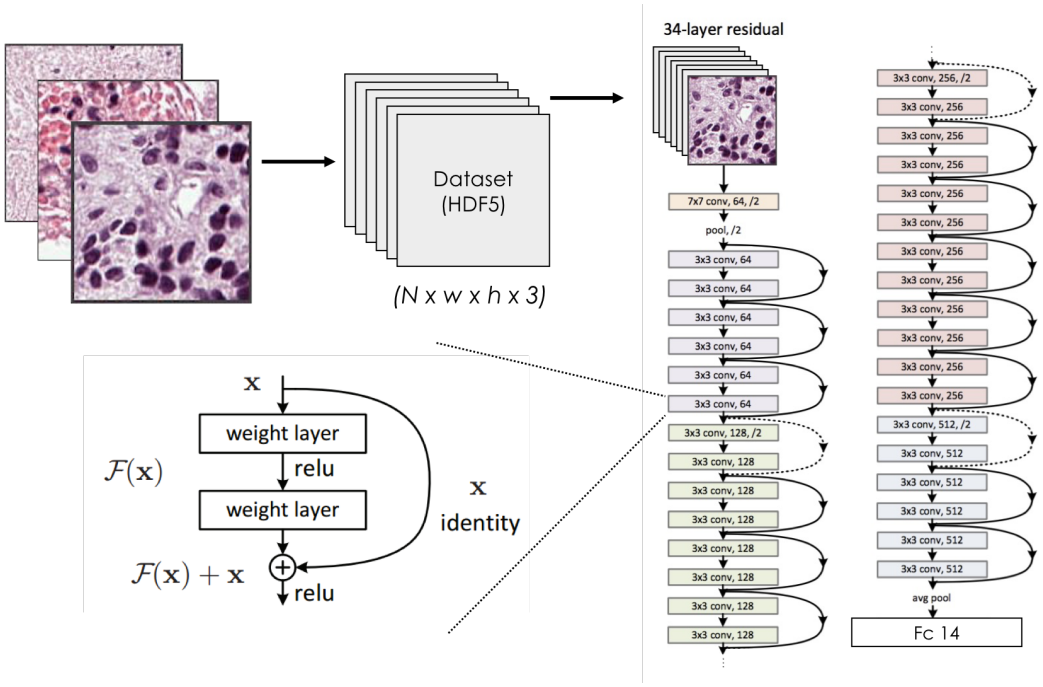


Figure 5.3: The ResNet50 model applied to histopathological image tiles. Whole slide images are tiled, labeled, and stored in a hierarchical data format on disk to reduce computational limits on random access memory availability. Tiles are stored as arrays composed of N tiles with width w and height h with three color channels in unsigned 8-bit integer format. The ResNet family of neural systems architecture incorporates skip connections between layers, as shown in the bottom left. These layers have been shown to improve the ability of the model to back-propagate error by reducing effects associated with the vanishing gradient problem.

in the metastatic setting as applied to the primary cancer setting. As shown in Figure 5.4, the previously trained model generally fails to remove undesired normal tissue from the primary setting, which is expected, as the first stage model was specifically trained to identify cancerous tumor strictly within liver tissue. Because this evaluation set is composed of primary sites we may reasonably expect that the model may be learning the morphological features associated with normal liver rather than morphological features associated with primary tumor. This finding necessitates the use of a primary-specific normal tissue filter model.

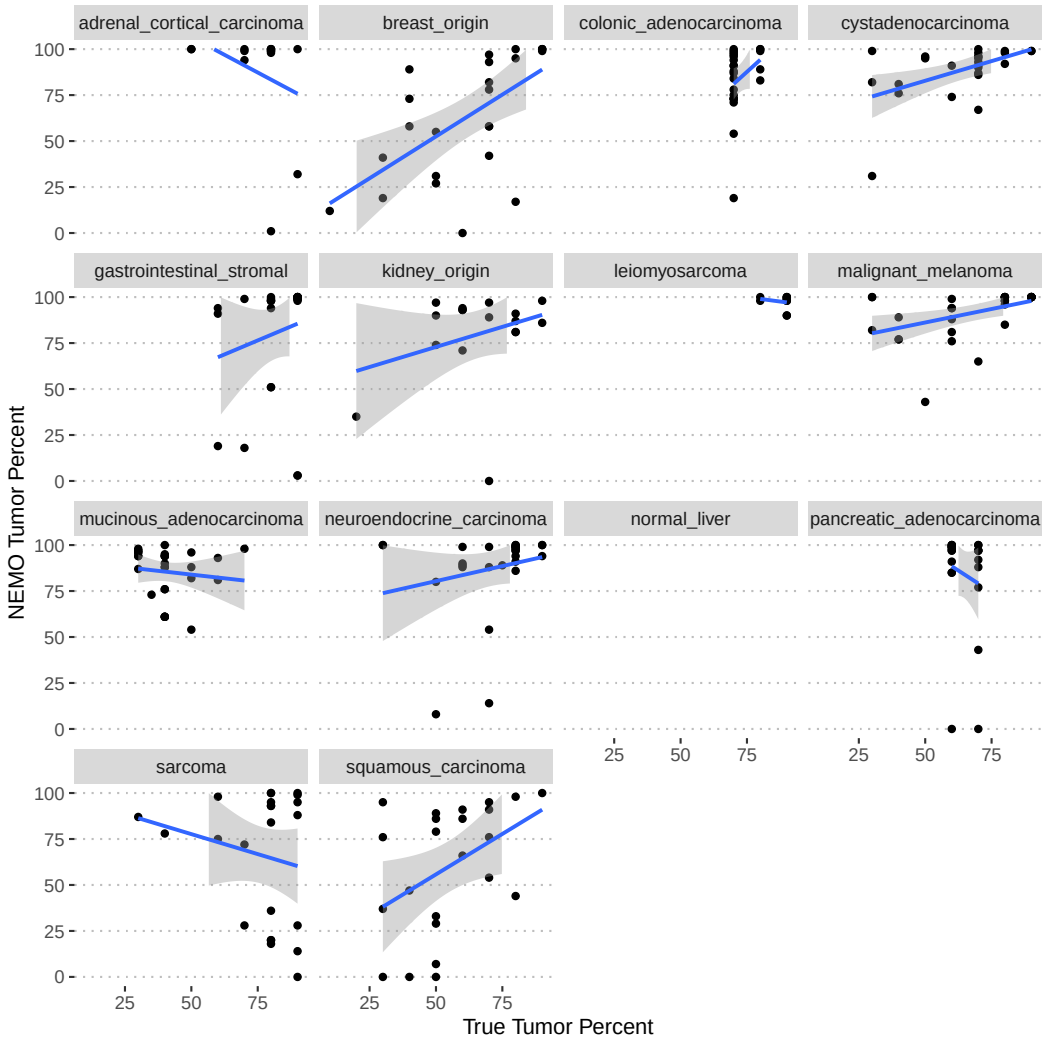


Figure 5.4: Percent tumor predicted by the previously-employed first stage NEMO model against the percent tumor annotation provided by the Knight BioLibrary.

5.3.2 Filtering Normal Primary Tissue

A new model was designed and trained to fit the classification labels provided by an expert pathologist. A ResNet50 model was trained to correctly predict whether a given tile was samples from within or external to the regions annotated as tumor by an expert pathologist. Annotations were generating using the QuPath[61] software tool for 78 whole slide images spanning all 14 distinct primary tissue types. Annotations were computationally extracted

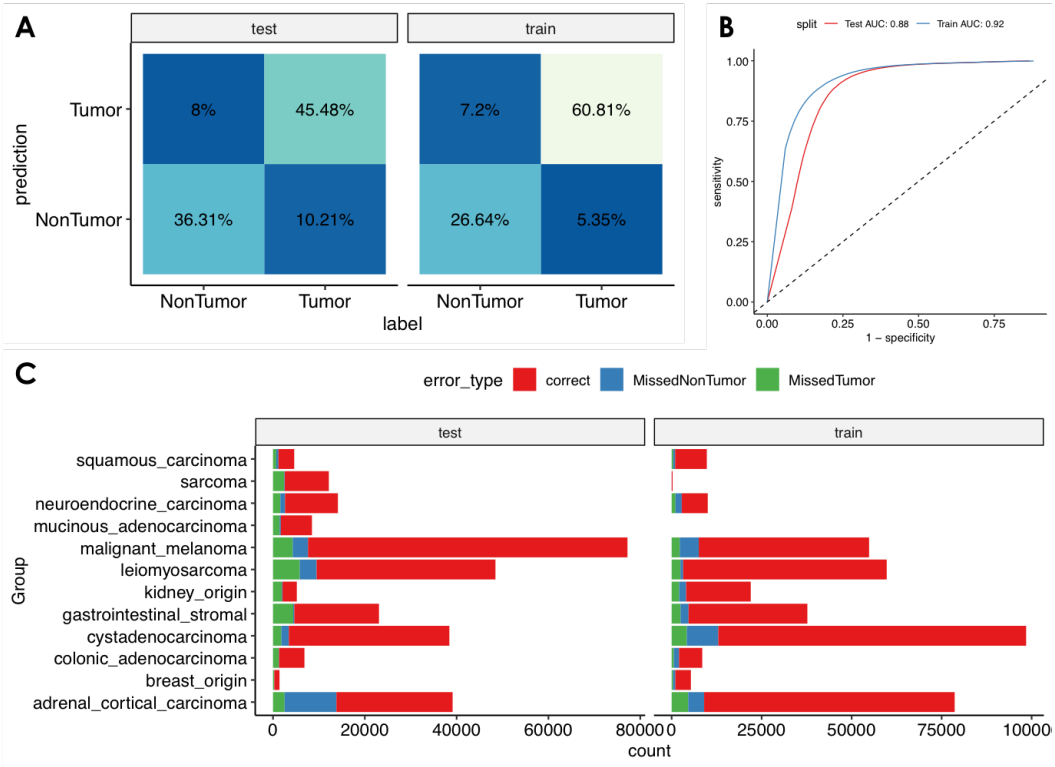


Figure 5.5: (A) Confusion matrices of a learned binary classification model to identify tumor tiles from normal tissue in primary cancers for training and held-out test sets. (B) Area under the receiver operator characteristics curve for both training and test sets illustrating an AUROC of 0.92 and 0.88 on the training and test sets, respectively. (C) Distribution of error types across the represented primary tumors under investigation illustrates relatively uniformly distributed error.

and employed to label each tile sampled from each of the 78 whole slide images as either belonging to a region annotated as tumor or not, in which case the tile is labeled non-tumor. Figure 5.5 illustrates the classification performance of the trained model as applied to a held-out test set. In general, the model performs well as evidenced by an area under the receiver operator characteristics curve of 0.88. The proportion of type I and type II errors were also shown to be generally evenly distributed across the fourteen classes, suggesting that no one class was responsible for model failure in either training or testing sets.

5.3.3 Primary-to-Primary Prediction

With both primary and metastatic datasets removed of normal tissue, second-stage models are required to correctly predict the tissue of origin of the remaining tiles deemed to contain tumor tissue. A ResNet50 model configured for 14-class output with softmax activation was trained on the primary dataset and evaluated on a held-out test set of whole slide images of primary cancer. The confusion matrix shown in Figure 5.6 illustrates good per-tile accuracy during training ($> 96\%$), though reduced generalization performance on the held-out testing dataset ($> 55\%$ accurate in 14-way classification).

A more nuanced understanding of the model's predictive ability is achievable if we loosen our success criteria away from requiring a strictly correct prediction. The distributions of rank to the correct choice is shown in Figure 5.7, which illustrates how far down the ordered set of predictions one must go to arrive at the correct prediction. For example, if a model's top choice is correct, then its correct rank is one, whereas if the correct prediction was the second highest confidence prediction, its correct rank would be two. While overall we can see that while the majority of samples were correctly predicted with rank 1, a sizable number of remaining samples appear to be within the top 5 or 6 ranks, suggesting that although the model might not make the correct prediction first, it may still enable a triage of potential selections based on their rank and confidence. Treating each predicted class as a binary classifier enables class-specific receiver operator characteristics analysis as shown in Figure 5.8, which illustrate good performance for each class individually for both tile and WSI predictions. Whole slide predictions are computed as the mean per-tile predictions across all tiles sampled from the respective whole slide, which supports the consistent observation that

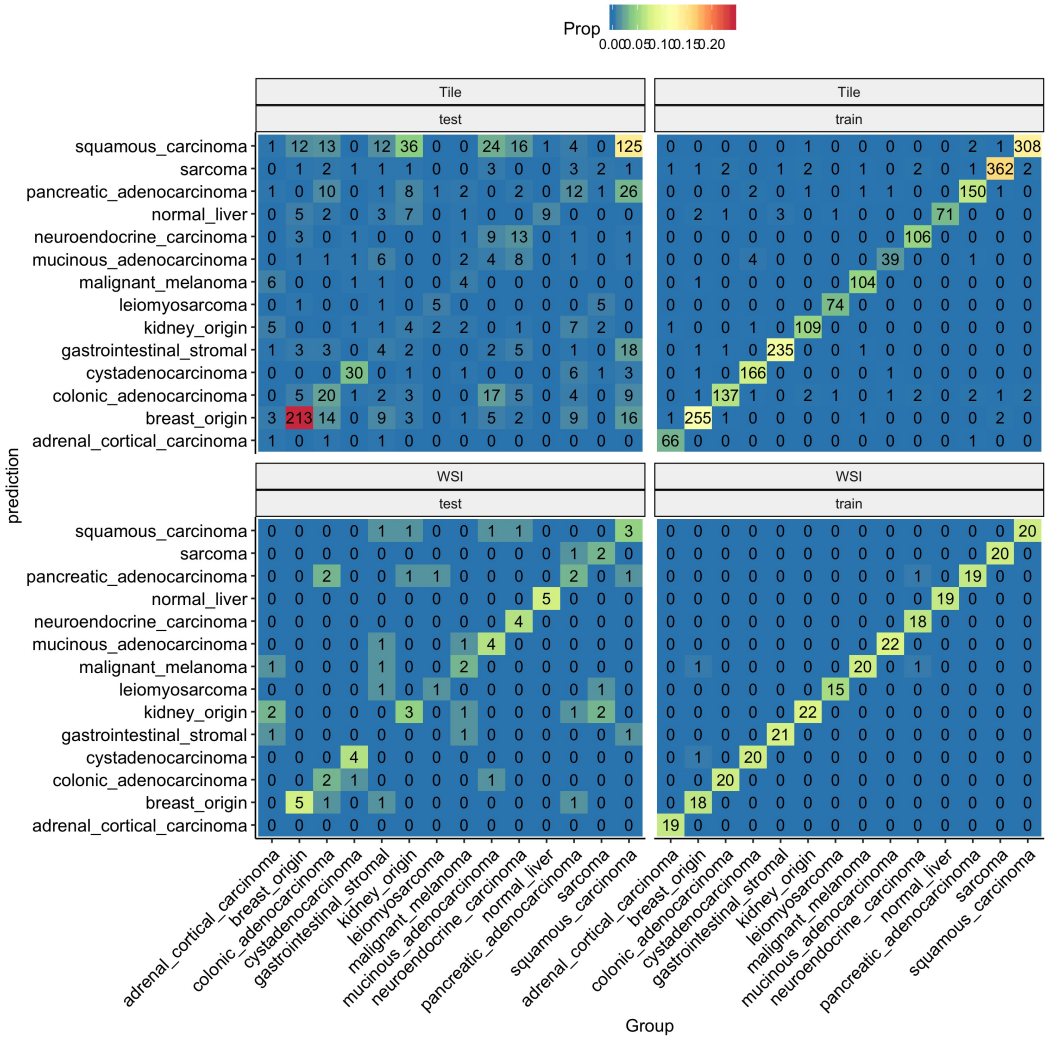


Figure 5.6: Distributions of the model's whole slide image confidence in the correct class, separated by whether the model was actually correct and shown for both training and testing datasets

the model's whole slide classification performance consistently out-performs the model's per-tile classification, as the approach smooths out intermittent tile-errors throughout the image. Because predictions are made on individual tiles independently, the spatial relationships among predictions of tissue type within the whole slide images is easily spatially-resolved. Figure 5.9 illustrates a single example of an incorrectly-classified whole slide image in which

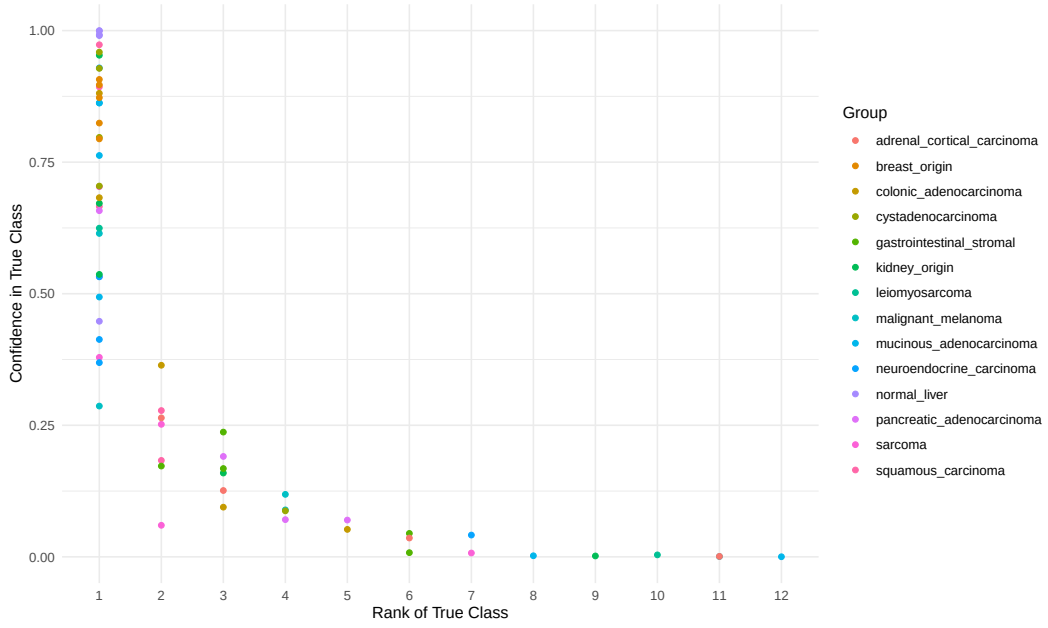


Figure 5.7: Distribution of the rank from true class. The x-axis reflects the rank-order of each slide's true class, so if a model's first, most-likely prediction is correct, that sample has a rank of true class of 1; if a model's second most likely choice was correct, then the rank of true class is 2. In general, the model is highly biased towards the correct class label. This supposes that although a model's most likely prediction may not be correct, considering multiple best guesses has a much higher likelihood of containing the correct option, similar to how a pathologist generates a differential diagnosis by considering various sources of evidence in support of various possible diagnoses of an unknown condition.

the correct whole slide label was a squamous carcinoma, but in which the model predicted the model to be a gastrointestinal stromal tumor. This perspective illustrates how although the model's first choice was incorrect, its second choice was correct with 35.2% and 18.3% confidence, respectively.

A deeper understanding of how and why the model fails to generate correct predictions for almost half of the samples may guide improvement to the training routine and refine our interpretation of model performance beyond strict accuracy. Figure 5.10 illustrates how the prediction confidence across the fourteen classes for both training and test set differ when the model is incorrect. By examining the samples that the model correctly classified

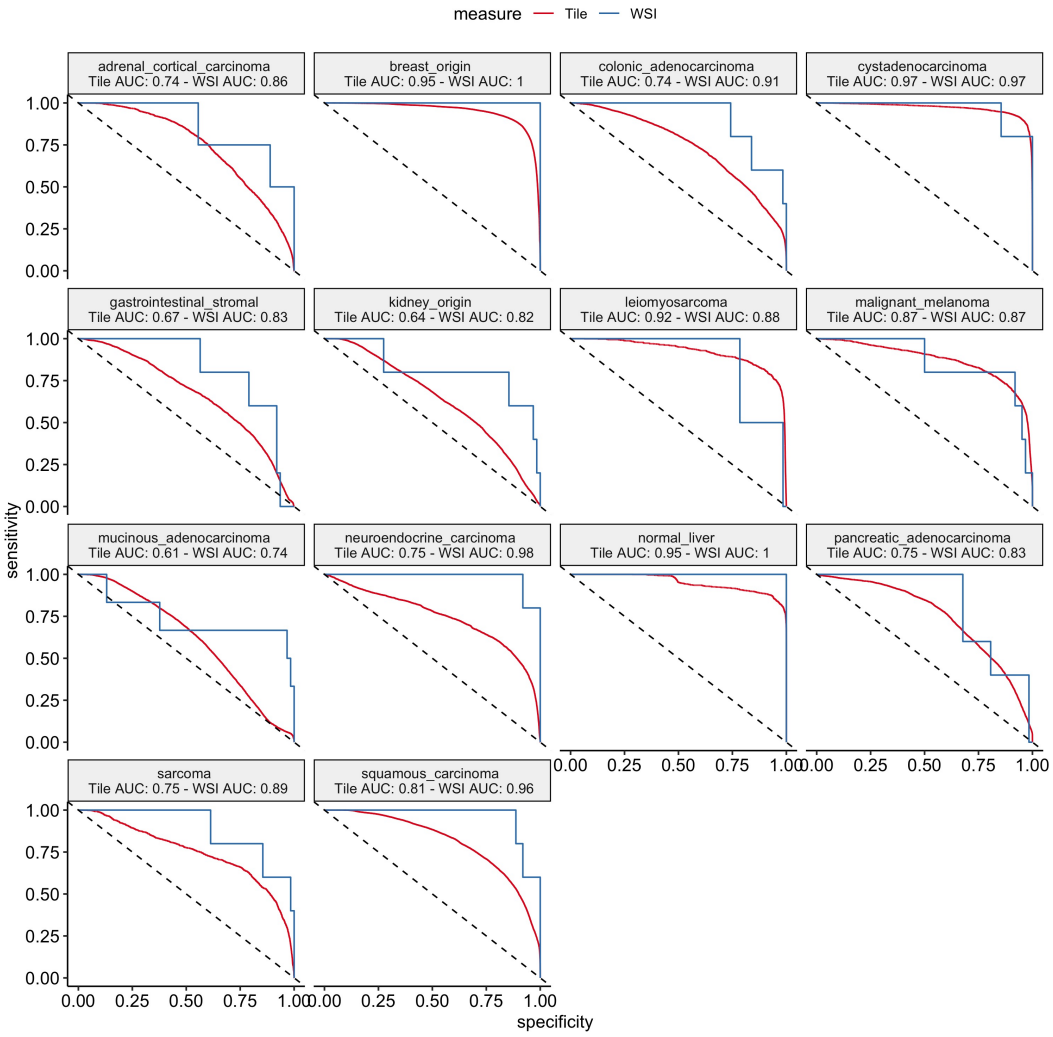


Figure 5.8: Area under the Receiver Operator Characteristics curves for tile and WSI predictions illustrates that in general predictions are more reliable at the whole slide level, which are made by computing the average predictions across all tiles in the whole slide image.

from those that were incorrectly classified, we observe a clear reduction in prediction for the correct class for incorrect samples, which is expected. However, we also observe a marked reduction in prediction confidence even for the samples that the model tends to get correct.

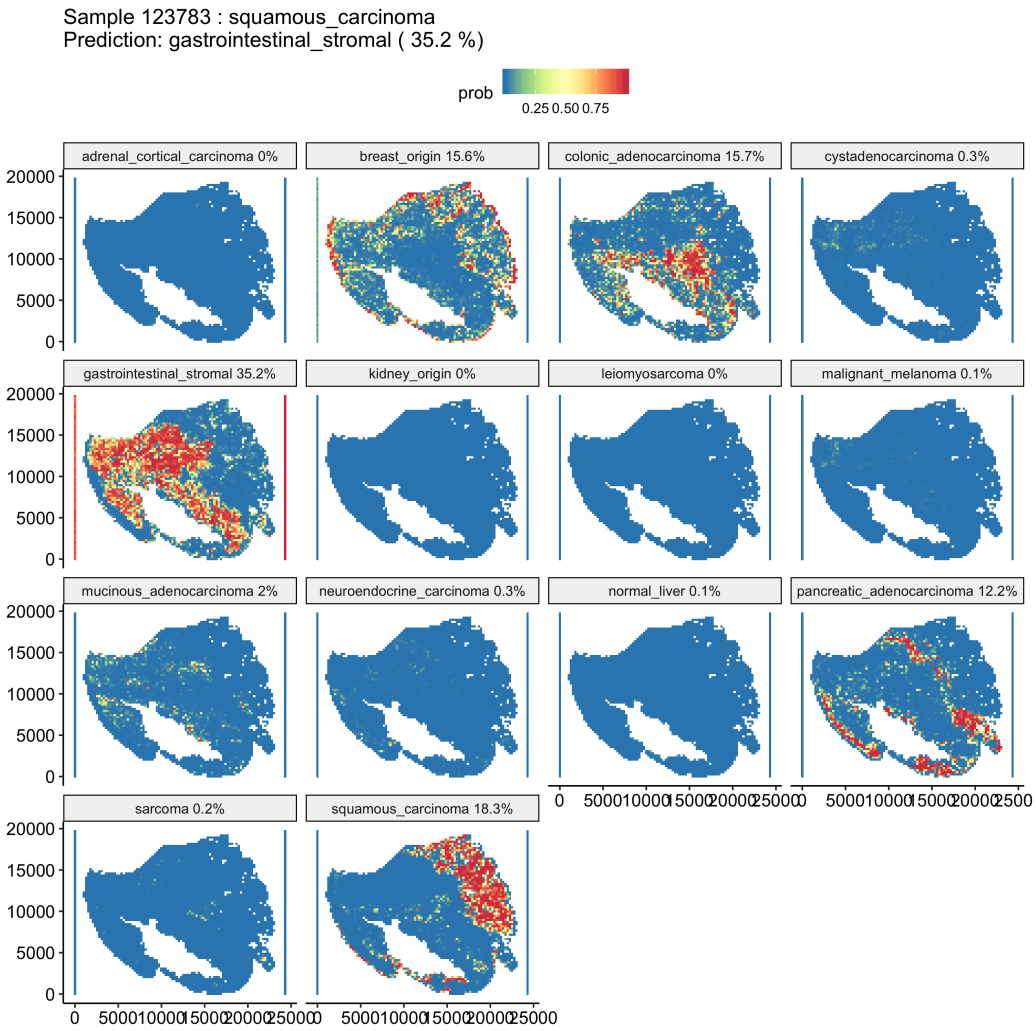


Figure 5.9: Example of spatially-resolved predictions from an incorrect classification of a squamous carcinoma incorrectly predicted to be a gastrointestinal stromal tumor. In this case, about half the tumor was correctly predicted to be a squamous carcinoma while the larger, other half of the image was incorrectly predicted to be a gastrointestinal stromal tumor.

5.3.4 Primary-to-Metastatic Classification

With both primary and metastatic samples filtered, a new classifier designed to predict the metastatic origin of secondary cancer is trained in three different manners. The first model is trained just on primary cancers as shown above. A second model is trained

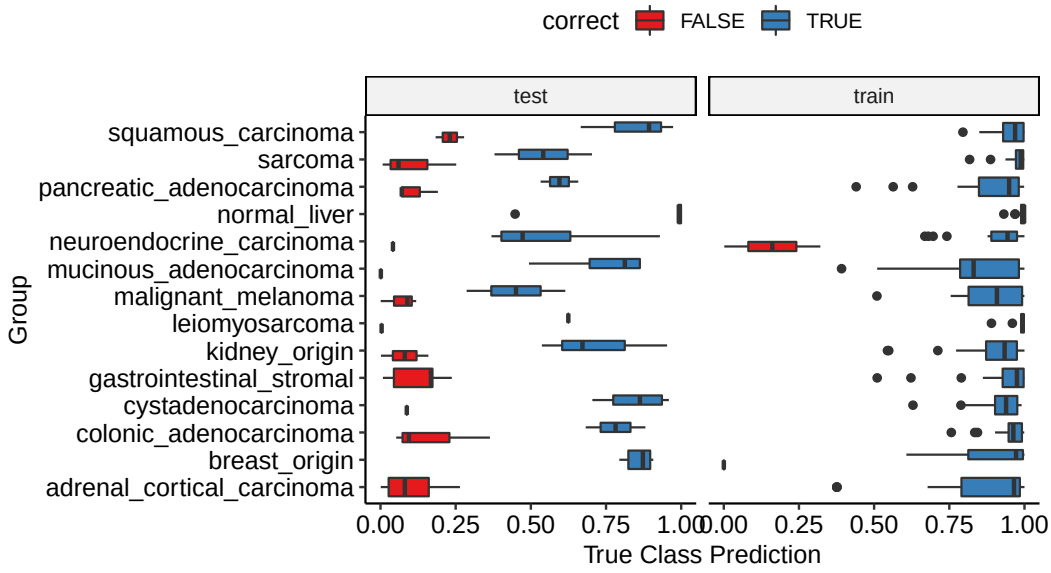


Figure 5.10: Distributions of model predictions of the correct class it was expected to choose, separated by both training and testing analysis and whether the model was correct.

on just metastatic cancers, and a third model is trained on primary cancers and then retrained on metastatic cancers. The third strategy describes a transfer learning approach, in which a model is first trained in one setting for one task - predicting the tissue of origin of primary cancer - and redeployed in a second setting on a different but related task, namely to predict the tissue of origin of metastatic cancer from whole slide images of liver metastases. In all three cases, an identical ResNet50 learning architecture was instantiated with identical Adam optimizers trained to minimize the cross-entropy between predictions and class labels. Data loaders were designed to generate batches of data with the batch size of 32 with the same data transforms that include flipping, rotating, and color scaling by saturation, brightness, and hue with class-balancing. To ensure constant training volume, the first two models are each trained for 10 epochs on their respective dataset, and the transfer learning model is trained for 5 epochs on each. After training is complete, each

model's state dictionary is saved to disc and reloaded upon evaluation of a held-out testing set composed of 159 metastatic whole slide images. Predictions are generated at per-tiles resolution, and whole slide predictions are computed as the average class prediction across all tiles in a whole slide image.

For each class, a receiver operator characteristics (ROC) curve is computed across the model's class-specific predictions for both tile and whole slide images for each of the three models as shown in Figure 5.11. These trends illustrate the quality of the model's predictions across the three training routines for both per-tile and per-WSI classification. In general, predictions at the whole slide level out-perform predictions made at per-tile level, which is not surprising as the simple whole slide averaging routine to generate whole slide predictions is capable of smoothing out local prediction inaccuracies and are thereby more robust to misclassification of tile subsets. Although in general, the model's performance is enhanced through the inclusion of pre-training on primary tumors, though this trend is not entirely consistent. For example, predictions of the metastatic origin of adrenal cortical carcinomas and gastrointestinal stromas appears to achieve best performance when trained on just primary tumor, while tumors originating in the kidney and squamous carcinomas appear to perform better when trained on just metastases. One plausible explanation for this observation might consider the inadvertent inclusion of surrounding normal tissue in either case, which may modulate the model's capacity to distinguish between the tumor types if the training labels are not applied to pure tumor tissue.

The quality of model training is also evident in examining the areas under the receiver operator characteristics (AUROC) curves, which provide a quantitative metric for model performance. The density distributions of AUROCs for each class shown in Figure 5.12(A)

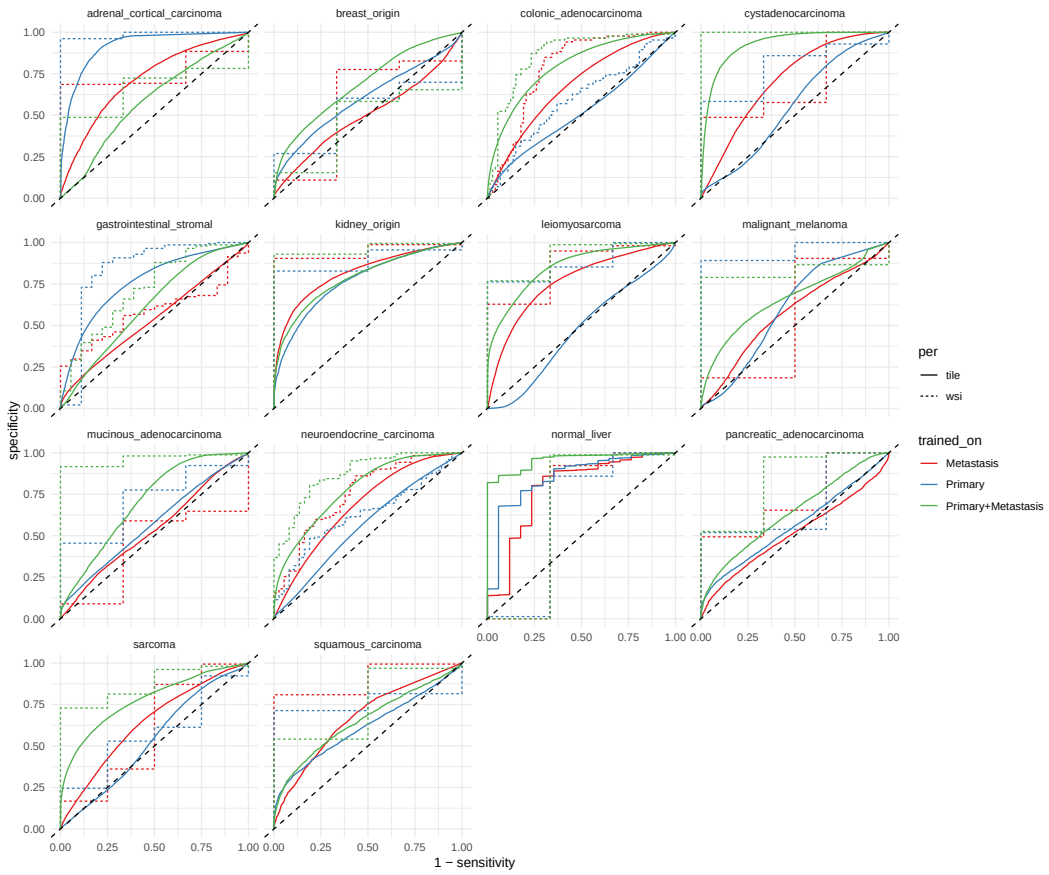


Figure 5.11: Receiver Operator Characteristic curves for the prediction of metastatic origin from three models trained on either primary tumor, metastatic tumor, or both. In general, combining metastatic and primary datasets contributes to superior classification performance using the primary-to-metastatic transfer learning framework described.

support the finding that in general pre-training on primary cancers enhancing model performance, which are broken down by specific class in Figure 5.12(B). The rank confidence plot in Figure 5.12(C), similar to Figure 5.7, supports similar findings that although the model's first choice may not be correct, the model's confidence is the correct class is generally higher than would be expected by chance. The per-tile confusion matrix shown in Figure 5.12(D) illustrates that the tile-classification task is generally more challenging for neuroendocrine carcinomas, colonic adenocarcinomas, and gastrointestinal stromas, which

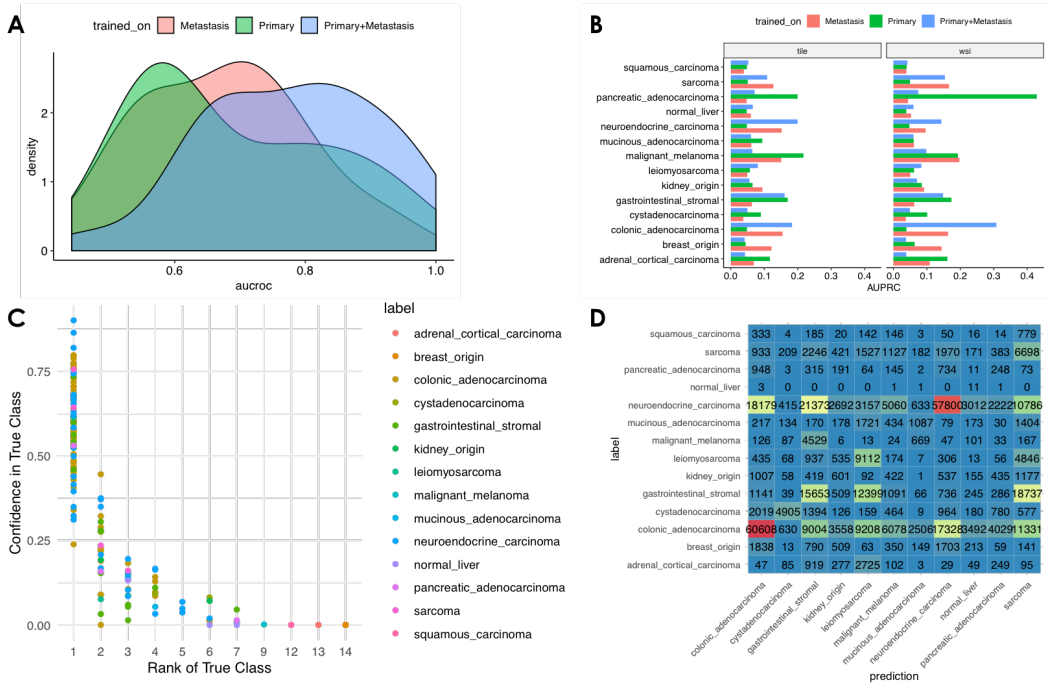


Figure 5.12: Whole slide classification of metastatic whole slide images. (A) Per-class AUROC distributions illustrate conferred benefit by incorporating both primary and metastatic samples in the metastatic origin prediction task. (B) Distributions of model performance across each of the predicted classes using each of the three described learning approaches. (C) Rank-confidence illustrates differential diagnosis potential, as the model's correct prediction tends to be non-uniformly distributed but instead biased towards the correct prediction. (D) Per-tile classification confusion matrix illustrating confounding variables in the colonic adenocarcinoma, neuroendocrine carcinoma, and gastrointestinal stroma classes, which also happen to be the most represented classes in the dataset.

are the most represented classes in the dataset.

5.3.5 Primary & Metastatic Feature Divergence

One plausible explanation for a model's failure to generalize between primary and metastatic settings might involve feature dissimilarity between images of a specific class of tissue between their primary and metastatic representations. We next evaluate whether inter-class divergence of image features between metastatic and primary tiles may be inferred to better explain discrepancies in how generalizable metastatic origin prediction is under the

hypothesis that cancer subtypes with greater discrepancy between primary and metastatic morphologies are less likely to be correctly classified by a model trained on primary cancers and deployed on metastatic cancers.

This approach trains an unsupervised variational auto-encoder on all tiles and evaluates how different cancer types differentially separate between primary and metastatic presentations. Facet tSNE embeddings of learned latent representations across the datasets are shown in Figure 5.13, illustrating the distribution densities of tiles belonging to a specific class colored according to whether they are from primary or metastatic samples. In principle, if tiles of a given cancer type look identical in both primary and metastatic images, then their distributions of tiles should perfectly overlap in this feature representation. This expectation supported by the tiles samples from normal liver in which tiles from both datasets almost perfectly overlap with one another. Other examples show similar but not complete overlap, particularly in the cases of colonic adenocarcinomas and pancreatic adenocarcinomas. However, other samples appear to be more widely distributed and non-overlapping, such as neuroendocrine carcinomas and squamous carcinomas. Some cases, such as leiomyosarcomas, appear strongly bimodal, suggesting that some tiles share similar features while others do not. The mean KL divergence across each of the learned features quantifies the statistical dissimilarity between primary and metastatic tiles, which are illustrated in Figure 5.14(A). Figure 5.14(B) illustrates that the correlation between mean KL divergence and the model's reported AUROC performance suggests a slight linear association with a correlation of 0.317, which may be interpretted as very slight evidence that the greater the divergence between primary and metastatic samples, the stronger the model's classification performance in the transfer learning setting. Figure 5.14(C) demonstrates

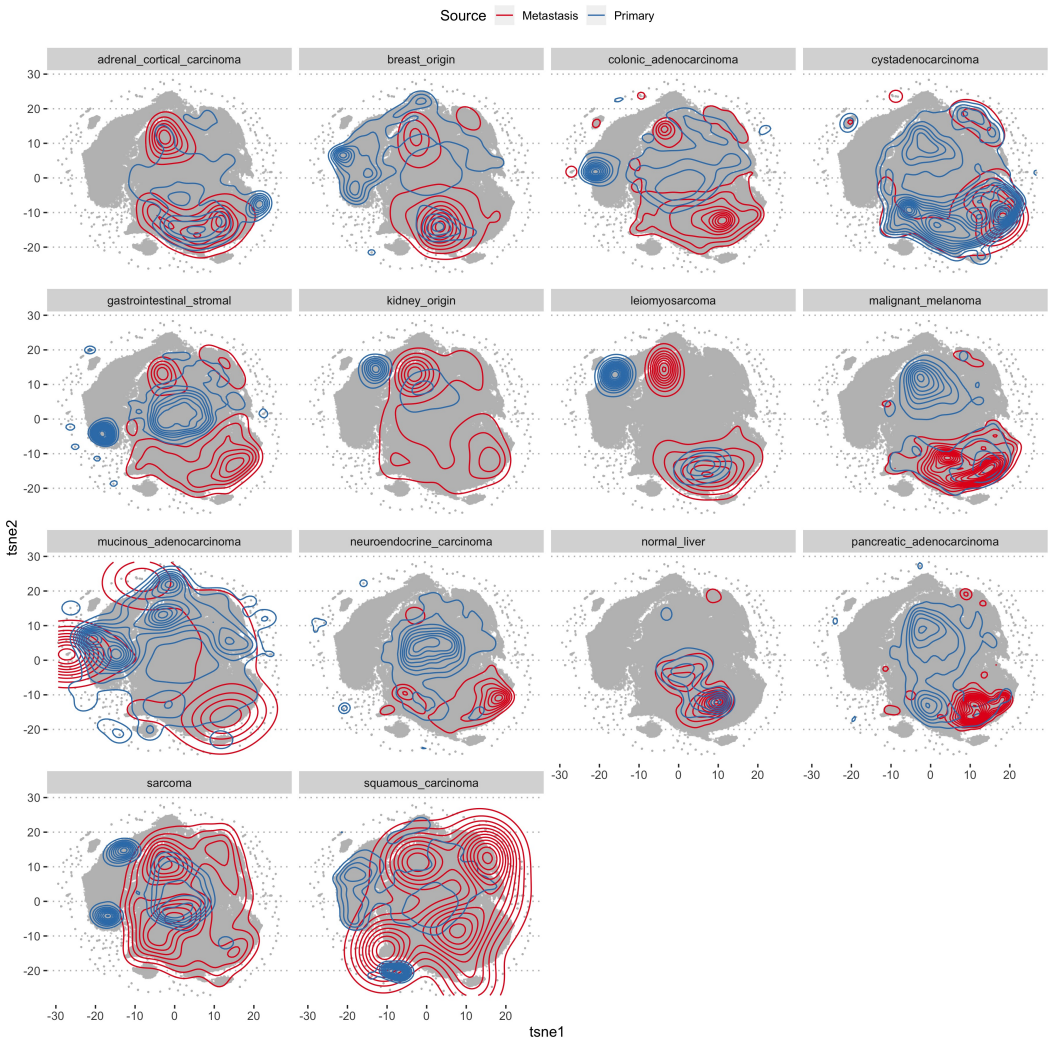


Figure 5.13: tSNE embedding of learned VAE features colored by both primary and metastatic samples

the mean divergence between pairs of tissue types, while 5.14(D) illustrates the divergence between primary and metastatic tiles across each of 64 learned VAE features. In general, these results support the visual interpretation taken from evaluating the distribution of tiles projected into two dimensions in the tSNE plots, showing that in general different cancer types exhibit distinct morphological feature representations.

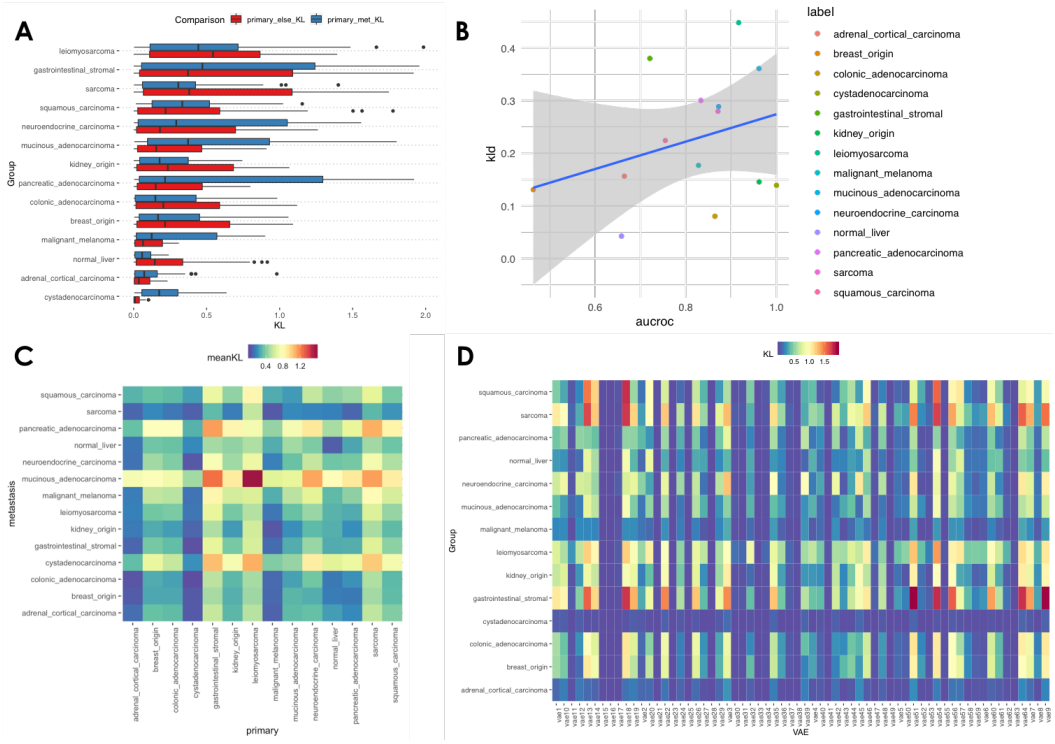


Figure 5.14: (A) Kullback-Leiber divergence across each of the learned VAE features illustrating separability between primary and metastatic samples measured both within and cross-class divergences. (B) Linear relationship between the KL divergence and the learned model's AUROC shows a surprising positive trend, suggesting that the more divergent the distributions, the better the chance of the learning model has to learn accurate classification. (C) Correlation matrix between primary and metastatic feature space representations from a single auto-encoder model. (D) Learned features associated with different classes of metastatic origin illustrates both class-specific and class-agnostic individual features.

5.4 Discussion

This work illustrates the importance of incorporating pre-training into histopathological classification, as pre-training on images of primary cancers confers benefit to a model trained to classify metastatic tumors according to their metastatic origin. This work opens many fruitful questions still unanswered, particularly with respect to the ability of these types of models to confer added benefit to a practicing pathologist tasked with inferring metastatic origin directly from histopathological whole slide images. This work has a number of limita-

tions that may limit the extensibility of its findings. In particular, this study was limited to 14 distinct classes of metastatic origin that were treated independently within the learning model. Future efforts may necessitate larger datasets with greater class-specific coverage to ensure robust ability to generalize both inter- and intra-class accuracy. This work also ignores shared latent features of tumor tissue that may be clinically relevant to rendering an accurate diagnosis of metastatic origin. Namely, this work ignores any other clinical feature of interest that may be relevant to this task, such as age, gender, medical history, and incidence of other disease phenotypes. This work also presents the clinical challenge of inferring metastatic origin in a simplified setting, when in practice a pathologist uses both other clinical covariate factors as well as obtainable results from axillary testing such as immunohistochemistry or genomic sequencing to guide their diagnostics.

The use of unsupervised feature extraction methods to infer feature divergence between primary and metastatic cancers also has a number of limitations. Like other efforts that incorporate these kinds of learning models, the learned feature space is subject to discrepancies, a lack of interpretability, and inconsistency in feature space embedding with similar input images. Although exploratory, it might be reasonable to use the relative dispersal of samples within a learned feature space to infer the heterogeneity or variability of intra-class samples so as to guide researchers in determining an adequate number of representative samples so as to cover an inferred feature space.

This work presents a number of future directions, in particular opening up the opportunity to explore clinical application of augmented inference of metastatic origin on a pathologist's classification performance. By measuring human accuracy both with and without the proposed model, clinical synergy between human expert and machine intelli-

gence may be quantified. This work may also present new opportunities to strategy patients based on computationally-inferred status of collected histopathological specimens. Retrospective analysis of the samples used in this study may enable a mechanism to train an auxiliary classifier to predict drug or treatment response from histopathological features of whole slide images. This line of research may contribute both to the design of therapeutic strategies, as well as to the design of investigative clinical trials seeking to stratify patience into competing study arms. Overall these results reinforce the importance of pre-training computer vision systems in digital pathology, as pre-training a network on primary cancers is shown to improve the ability of the learning model to infer the origin of secondary metastatic cancer.

Chapter 6

Conclusion

> I think... that you have to have more than just a machine.
>
> All the problems of the world could be settled easily, if men were only willing to think. The trouble is that men very often resort to all sorts of devices in order not to think, because thinking is such hard work.
>
> *Thomas J. Watson,*
> *Chairman & CEO, IBM (1914–1956)*

This work has investigated computational strategies to employ deep learning systems architecture for histopathological image analysis in a clinical setting. This chapter summarizes the limitations broadly applicable to this approach, as well as future research directions in the field of digital pathology & artificial intelligence.

6.1 Limitations

The artificial neural networks described in this work vary in their type, complexity, and intended purpose, yet their use has been motivated by the clinical need for faster, cheaper, and more reliable methods of interrogating tumor biology. This work, which largely revolves around metastatic cancer of the liver, is both promising yet limited. Although the methods

utilized in this work are not necessarily new, this work was not intended to introduce an incremental improvement in a computer vision task, nor was it advance underlying knowledge of cancer biology. Rather, this work is interdisciplinary by its very nature, and illustrates how clever application of existing tools to important problems can in fact yield interesting and surprising insights as by-products of their application. Cutting edge research in digital pathology tends to incorporate many more hundreds or thousands of whole slide images than what was included in these studies. Considering the significant heterogeneity of many cancer types, the datasets used to model computer vision and pattern recognitions systems should then be trained on equally diverse data. The collection, storage, and management of these large datasets presents a significant challenge, and though this work has benefited from an accessible data repository, it is not shown whether the models presented would fairly generalize to images scanned and retained by other slide repositories. As large consortia continue to interconnect data and images in support of the research community, additional efforts must be made to ensure that learning strategies validate both within and external to a host repository.

Some of this work was initially intended to serve as a means of augmenting a human pathologist's capacity to interrogate whole slide images. However, these results do not fully close that gap. More comprehensive evaluation of both the learned models as well as their effects on human operators might be necessary to achieve acceptance into clinical practice.

Future work in digital pathology and artificial intelligence appears to encounter a fork in philosophical approach to building sophisticated learning systems, forcing practitioners to choose whether an artificial intelligent solution should be specialized to precise singular tasks, or generalizable as human pathologists are trained to be. Both approaches have their

own limitations, but because this work has not pursued general-purpose histopathological feature learning, limitations of the specialized approaches are relevant. Specialized systems may do well on clearly defined tasks, but they pose a significant additional problem in determining how best to integrate specialized networks to fulfill more general problem solving. For example, some of this work illustrated multi-stage approaches where one specialized network might filter out surrounding stroma tissue, while a second more specialized network was trained to classify the remaining tissue. In this case, it is relatively straight forward to integrate these models in series, one performing its task after the other. But humans are more complex in their decision-making which has not been modeled by the approaches described here, presenting a new limitation in the capacity for any of the learned models presented here to work in symphony with other computer vision systems.

Relatedly, general computer vision systems are generally designed to process input of common shape and size such that the number of trainable parameters is held fixed across an entire dataset. In this work, design decisions regarding the size of the samples tiles were made largely based on established practices in the field. While this approach is sufficient to achieve certain results, it ignores the function of the human vision system and the manner by which trained pathologists examine sections of tissue, in which liberal use of lateral movement and zoom create highly dynamic inspection processes during which the pathologist considers complex spatial information at variable scale. To date, few systems are designed to mimic this behavior, and so are limited in their ability to leverage spatial information that requires far-field context to meaningfully interpret.

6.2 Future Research Directions

Upon setting out upon this work, three things were noted of relevant import: pathology is in many ways a gold-standard for cancer diagnosis and, therefore, cancer treatment; inter-operator agreement between pathologists is often surprisingly poor; data availability in digital pathology is quite good; and deep-learning methods are powerful enough to likely contribute to resolving certain discrepancies in the image analysis work-flow at the core of pathological practice.

The vast majority of the work that went into designing, developing, and verifying computational pipelines to execute the research presented is not emphasized, nor indeed should it be; this dissertation is primarily motivated to share new science and knowledge in the field of biomedical engineering. However, the reader might consider how the tools and methods employed in these specific studies might easily extend themselves to very similar problems in very different domains. Classifying bits of tissue as one type or another is not unique to cancer pathology, nor is the need to identifying distinct tissue or cells. The contributions of this work are therefore two-fold: first, the science that emerges from the motivating questions; and second, a tool platform that is easily reconfigured to begin answering many more questions in a much shorter amount of time than what was required to prepare this dissertation.

AI for Differential Diagnoses

A current future direction under considering considers secondary utility for a model's differential prediction of metastatic origin. Although a single model making a single prediction is

Figure 6.1: Incorporating IHC panel recommendations into a learning model's differential diagnostic predictions retains an expert pathologist's decision-making and axillary molecular testing methods.

sometimes correct, in practice clinicians often rely on *differential diagnoses* which typically rank-order possible underlying conditions given various bits of medically-relevant evidence. A neural network's probabilistic output can be interpreted as a differential diagnosis, since it enables an observer to rank order the model's predictions based on the model's confidence across available choices. In clinical practice, a pathologist may leverage relative uncertainty to order a confirmatory IHC panel designed to differentiate between potential diagnoses. On-going work led by Hassan Ghani has integrated a knowledge base of discriminatory IHC stains into a web-based queriable database that tabulates and ranks stains according to their discriminatory power, such as teh example shown in Figure 6.1. The models presented in this work may be extensible in that manner, such that the model's most likely candidate choices may be employed to optimally select distinguishing IHC stains without the intervention of a biased clinician.

Electron Microscopy Annotation

While some of the work presented herein has commented on the data size of images used for training deep learning models, the volume of data from immunohistochemistry pales in comparison to data volumes generated by scanning electron microscopy. Such ultra-high-definition measurement systems enable visualization of tissue in a single channel at 4nm resolution and in three spatial dimensions, offering unprecedented resolution into the structural composition of tumor tissue. The DeepHAT system described in chapter 4 is already under exploratory investigation to evaluate whether annotations of these such images may be accelerated using clever feature space projection approaches.

Predicting Drug Response

Significant clinical need exists for increased availability of useful drugs with accompanying actionable bio-markers, and one particularly rich area of open questions lies in the intersection of paired patient histopathology and clinical trial outcome. One can imagine stratifying a patient population based on their response to investigative therapy and training a classification network to predict which patients are more or less likely to respond. If such a model can be trained to generalize well to new patients, secondary opportunity lies in the careful interrogation of the model's reasoning to infer new bio-markers that themselves might be predictive of patient response. However, if interpreted reasoning eludes the observer while a neural network model continues to generalize well, new drugs may be accompanied by the computer vision system themselves to serve as the bio-marker retrieval system, rather than relying on a pathologist's determination.

Human Factors Engineering

Current scientific literature contains an abundance of evidence that suggests that the likelihood that computer vision systems will soon routinely meet or exceed human capability are on the horizon. Yet despite published models or datasets, academia has largely ignored the value of human factors engineering to identify mechanisms that might most effectively incorporate artificial intelligence learning systems into clinical practice. While the input of pathologists is often used to explicitly label data or confirm a model's performance, rarely are the thoughts and opinions of pathologists regarding how they might best prefer to interface with AI incorporated into published design decision making. A related line of inquiry should investigate down-stream effects associated with over-reliance on AI-based methods. A hypothetical experiment might be designed to evaluate the role of an AI-system on performing some task, evaluate the role of a human expert at that task, and evaluate the role of the human plus the AI-system, with the hypothesis being that a synergistic effect might be measurable when combining the relative strenghts of the two actors. However, an important twist might then investigate the effect of that synergy as the AI system is gradually impeded, or handicapped, without the human operator's knowledge. Do humans tend of forsake effort when machines appear competent? And, if so, can we measure the degree of error expected if a machine later begins to fail at its task, and can we anticipate whether a human will intervene and take corrective action? In clinical medicine, the human cost of an over-reliance on artificial intelligence is unknown; the human factors engineering of AI integration is ripe for exploration and innovation.

Extension to Multiplexed Imaging

Although typical microscopy is limited by the bands of the electromagnetic spectrum to which human eyes are tuned, fluorescent labels enable wide-field view of tissue make-up with protein-specific channels of information, enabling generation of imaging data with far greater channel depth than the typical red, blue, and green channels that can recapitulate a significant portion of the electromagnetic spectrum human eyes are tuned to. Emerging technologies including cyclic immunofluorescence, CODEX, and mIHC are all means of acquiring two-dimensional images with varying independent axes of depth not from color, but from the presence of molecules and proteins. Though much of the digital pathology work presented in this thesis has relied on H&E images, characteristically occupying mostly just red (eosin) and blue (hematoxylin) color channels, the methods described herein are largely indifferent to the channel depth of data object, enabling translation of these methods to these exciting emerging technologies. Neural network architecture, and the convolutional layers from which modern computer vision systems are built, are highly amenable to tensor inputs of arbitrary dimensionality, simply swapping the depth channel which is nominally three for color images, to N for however many independent measurements are able to be captured with a multiplexed imaging platform.

Imaging-Omics Integration

As the depth of information extractable from whole tissue continues to increase, driven in part by the emergence of multiplexed imaging technology, and as single-cell segmentations become more robust, to integration of rich imaging data with rich genomics data presents a significant opportunity to uncover causal and associated relationships between these two

great pillars of biomedical data. Deep learning methods of contributed to significant progress in cellular segmentation, particularly in multiplexed imaging domains where cytoplasmic-specific markers provide convenient delineations between cell bodies. These approaches are not generating large, spatially-resolved datasets of proteomic signatures at single-cell resolution, in many ways similar to popular single-cell genomic measurements techniques, yet retain spatial context of measurements. Figure 6.2 illustrates the core principle behind a proposed method to be published in this year's 59th annual Conference on Decision and Control in which translation between two measurements of the same unit of analysis, in this case the cell, may be learned even in the absence of unpaired measurements. In this work, a self-supervised domain translation model may be learned by enforcing a cycle-consistent reconstruction penalty as a sample from one domain is cast into a mutually-shared feature space in some other domain before being translated back again. Although this work is still preliminary, it holds great promise for elucidating biological mechanisms that are currently difficult to interrogate given absence of paired multiplexed data sources.

6.3 Dei Ex Machinae

The human brain is thought to possess on the order of one quadrillion synaptic connections between pairs of approximately ninety billion neurons. This astounding complexity is without comparison in the known universe, and dwarfs the computational complexity of even the most advanced artificial neural network systems. If we were to draw an analogy between the number of trainable parameters in an artificial neural network and the number of synapses within the brains of living creatures, then we must resign ourselves to examining

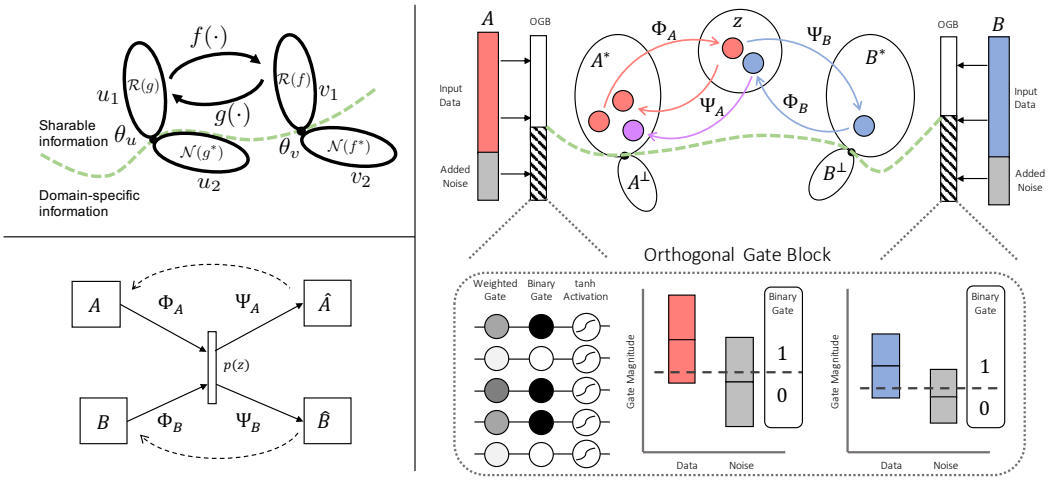


Figure 6.2: Integrating single-cell imaging and omics data remains a great challenge in systems biology. The proposed XAE learning system architecture is designed to learn cross-domain translation from unpaired datasets.

the so-called "lower-order" organisms of our planet to find familiar kin. The thirteen million trainable parameters of the popular ResNet50 model would scantly supersede the nearly ten million synaptic connections of the common fruit fly, *Drosophila Melanogaster*. Though the likelihood that a fruit fly could learn to interpret histological slides is no greater than the likelihood that an artificial neural network might grow wings and learn to fly, the parallels of the scope and scale of the computational complexity of each motivate an appreciation for the evolutionary future of artificial learning systems, as they gain rapidly upon the slower, less coordinated evolution of biological thinking machines.

The populist hype surrounding deep learning is no accident, nor is it without justification. The performance of these systems, demonstrated in this work and in the works of many others, are an encouraging sign that artificial neural networks are capable of achieving remarkable results, even with respect to well-trained human experts. In the eyes of a practitioner, their success is due in no small part to their elegance, accessibility, and flexibility.

But in the eyes of the skeptic, concerns are amplified by both researchers and the public as the field of artificial intelligence continues to blaze a trail of public failures alongside a broad avenue of success. Despite failures, setbacks, and limitations, the opportunity nevertheless continues to attract investments in the billions of dollars, pushing valuations of deep learning-centric enterprises to the highest reaches of venture capital.

Medical science is accelerating at a rate beyond the ability of a human mind to keep pace. Yet machine learning systems, such as those described in this document, though operating at a level far beneath the capacities of a human mind, are capable of collecting, disseminating, and processing data in vast quantities and at near constant rate. The future of these tools to improve diagnostics in health care at home and around the globe is, in the humble opinion of the author, inevitable. The role of machine intelligence systems in understanding, characterizing, and diagnosing disease is not yet well-formulated. To trust a computer to make decisions or infer causality of disease inherently places the well-being of fellow humans into the nebulous decision-making processes within a computer system. But just as a self-driving automobile does not have to have a 100% success rate to be better than a human driver, so too do machine learning models need only surpass human decision makers in narrowly-defined tasks to provide value to health care delivery. If nothing else but speed and reproducibility, systems like those presented in this work have measurably valuable qualities that make them potent partners in the battle against disease.

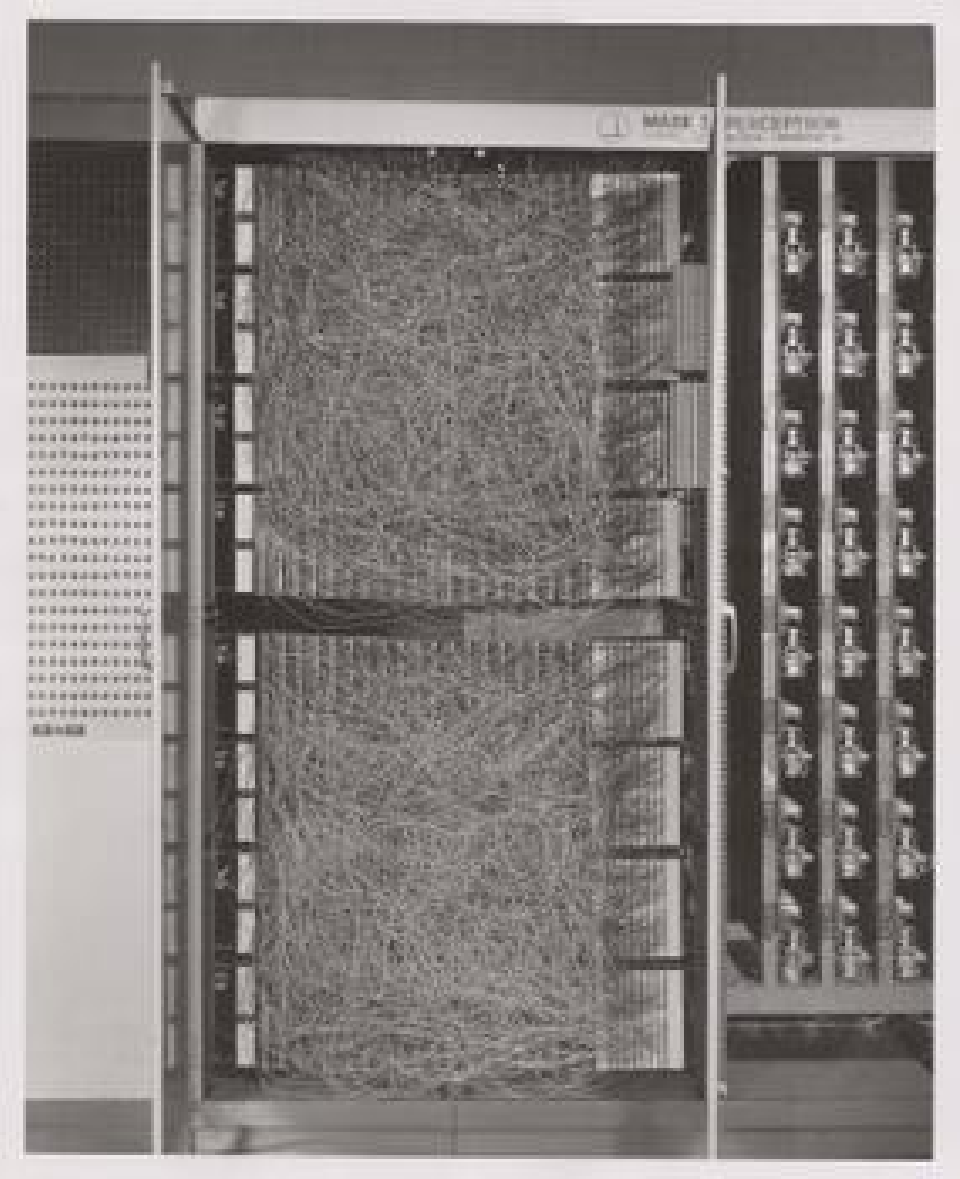
To Not Replace...

At the Frontiers of AI in Medicine Symposium held at Stanford University in September, 2019, Professor Fei-Fei Li, conference chairperson and prominent leader of the AI field in

computer vision, introduced the symposium with a statement that she believes that we, as a community, must hold in paramount importance: *That the objective is not to replace, but to augment.*

It is true that certain actors in the world are motivated to replace doctors and care-givers by unfeeling robotic analogs deemed superior by virtue of their capacity to ingest and process information. A certain self-righteousness and hubris within the technocratic strata might claim to be on the cusp on throwing out the human system upon which medicine is based in favor on some type of impartial thinking machine. The philosophical divide between believers and doubters of AI is widening, but a middle ground reasonably holds that machines are no more capable of replacing doctors than humans are of replacing the microchip.

Organizations across all strata of biology, from cells and tissue through societies and ecologies, rely on the competitive advantage of their constituent member units to sustain life. This work has sought to bring together the unique capabilities of human and machine towards the betterment of our own species, and refinement of our machine counterparts, through the study of cancer. The natural symbiosis between human and machine, each reinforcing the distinct advantages of the other, will continue to offer profound contributions to the medical sciences, extend the well-being of our species, augment our role as care-givers, and facilitate the continued democratization of medicine to every member of our species and to every corner of the globe. May thinking machines of the future herald good and sustainable health for all humankind.



The Mark I, the first implementation of the perceptron algorithm (1958)

Bibliography

[1] Henrica M J Werner, Gordon B Mills, and Prahlad T Ram. Cancer Systems Biology : a peek into the future of patient care? *Nature Publishing Group*, 11(March):167–176, 2014.

[2] Eung-sam Kim, Eun Hyun Ahn, Euiheon Chung, and Deok-ho Kim. Recent Advances in Nanobiotechnology and High-Throughput Molecular Techniques for Systems Biomedicine. *Molecular Cells*, 36:477–484, 2013.

[3] Jorrit J Hornberg, Frank J Bruggeman, Hans V Westerhoff, and Jan Lankelma. Cancer : A Systems Biology disease. *BioSystems*, 83:81–90, 2006.

[4] Douglas Hanahan and Robert A Weinberg. The Hallmarks of Cancer. *Cell*, 100:57–70, 2000.

[5] Douglas Hanahan and Robert A Weinberg. Review Hallmarks of Cancer : The Next Generation. *Cell*, 144(5):646–674, 2011.

[6] Joe Gray. Spatial systems biology of cancer. *Molecular Cancer Research*, 14(2 Supplement):IA06 LP – IA06, feb 2016.

[7] Sjoerd Rodenhuis, Marcus L. Van De Wetering, Wolter J MooI, Siegina G. Evers, Nico Zandwijk, and Johannes L Bos. Mutational Activation of the K-RAS Oncogene. *New England Journal of Medicine*, 317(15):929–935, 1987.

[8] D. P. Lane and S. Benchimol. p53: Oncogene or anti-oncogene? *Genes and Development*, 4(1):1–8, 1990.

[9] Peter Carmeliet and Rakesh K Jain. Angiogenesis in cancer and other diseases. *Nature*, 407:249–257, 2000.

[10] John Maciejowski and Titia De Lange. Telomeres in cancer: Tumour suppression and genome instability. *Nature Reviews Molecular Cell Biology*, 18(3):175–186, 2017.

[11] Chun Wen Cheng, Pei Ei Wu, Jyh Cherng Yu, Chiun Sheng Huang, Chung Tai Yue, Cheng Wen Wu, and Chen Yang Shen. Mechanisms of inactivation of E-cadherin in breast carcinoma: Modification of the two-hit hypothesis of tumor suppressor gene. *Oncogene*, 20(29):3814–3823, 2001.

[12] Gaorav P Gupta and Joan Massagué. Cancer Metastasis: Building a Framework. *Cell*, 127:679–695, 2006.

[13] Patricia S Steeg. Targeting metastasis. *Nature Publishing Group*, 16(4):201–218, 2016.

[14] Johanna A Joyce and Jeffrey W Pollard. Microenvironmental regulation of metastasis. *Nature Reviews Cancer*, 9(April):239–252, 2009.

[15] Denis Wirtz, Konstantinos Konstantopoulos, and Peter C. Searson. The physics of cancer: The role of physical interactions and mechanical forces in metastasis. *Nature Reviews Cancer*, 11(7):512–522, 2011.

[16] James E. Talmadge and Isaiah J. Fidler. Aacr centennial series: The biology of cancer metastasis: Historical perspective. *Cancer Research*, 70(14):5649–5669, 2010.

[17] Sean G Megason and Scott E Fraser. Imaging in Systems Biology. *Cell*, 130:784–795, 2007.

[18] Armen R Kherlopian, Ting Song, Qi Duan, Mathew A Neimark, Ming J Po, John K Gohagan, and Andrew F Laine. A review of imaging techniques for systems biology. *BMC Systems Biology*, 2(74):1–18, 2008.

[19] Laura Orlando, Giuseppe Viale, Emilio Bria, Eufemia Stefania, Isabella Sperduti, Luisa Carbognin, Paola Schiavone, Annamaria Quaranta, Palma Fedele, Chiara Caliolo, Nicola Calvani, Mario Criscuolo, and Saverio Cinieri. Discordance in pathology report after central pathology review: Implications for breast cancer adjuvant treatment. *The Breast*, 30:151–155, 2016.

[20] Alexander J J Smits, J Alain Kummer, Peter C De Bruin, Mijke Bol, Jan G Van Den Tweel, Kees A Seldenrijk, Stefan M Willems, G Johan A Offerhaus, Roel A De Weger, Paul J Van Diest, and Aryan Vink. The estimation of tumor cell percentage for molecular testing by pathologists is not accurate. *Modern Pathology*, 27:168–174, 2014.

[21] Paula A Rodriguez-urrego, Angel M Cronin, Hikmat A Al-ahmadie, Victor E Reuter, and Samson W Fine. Interobserver and intraobserver reproducibility in digital and routine microscopic assessment of prostate needle biopsies. *Human Pathology*, 42(1):68–74, 2011.

[22] Jon Griffin, Darren Treanor, J Griffin, and D Treanor. Digital pathology in clinical use : where are we now and what is holding us back ? *Histopathology*, 70:134–145, 2017.

[23] Takahiro Tsujikawa, Sushil Kumar, Rohan N Borkar, Joe W Gray, Paul W Flint, Lisa M Coussens, Takahiro Tsujikawa, Sushil Kumar, Rohan N Borkar, Vahid Azimi, Guillaume Thibault, and Young Hwan Chang. Quantitative Multiplex Immunohistochemistry Complexity Associated with Poor Prognosis Resource Quantitative Multiplex Immunohistochemistry Reveals Myeloid-Inflamed Tumor-Immune Complexity Associated with Poor Prognosis. *CellReports*, 19(1):203–217, 2017.

[24] Edward C Stack, Chichung Wang, Kristin A Roman, and Clifford C Hoyt. Multiplexed immunohistochemistry , imaging , and quantitation : A review , with an

assessment of Tyramide signal amplification , multispectral imaging and multiplex analysis. *Methods*, 70(1):46–58, 2014.

[25] Jia-ren Lin, Mohammad Fallahi-sichani, and Peter K Sorger. Highly multiplexed imaging of single cells using a high-throughput cyclic immunofluorescence method Jia-Ren. *Nature Communications*, pages 1–7, 2015.

[26] Charlotte Giesen, Hao A O Wang, Denis Schapiro, Nevena Zivanovic, Andrea Jacobs, Bodo Hattendorf, Peter J Schüffler, Daniel Grolimund, Joachim M Buhmann, Simone Brandt, Zsuzsanna Varga, Peter J Wild, Detlef Günther, and Bernd Bodenmiller. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature Methods*, 11(4), 2014.

[27] Yury Goltsev, Nikolay Samusik, Julia Kennedy-darling, Gustavo Vazquez, Sarah Black, Garry P Nolan, Yury Goltsev, Nikolay Samusik, Julia Kennedy-darling, Salil Bhate, Matthew Hale, and Gustavo Vazquez. Deep Profiling of Mouse Splenic Architecture with Resource Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell*, 174(4):968–981.e15, 2018.

[28] Andrew H Fischer, Kenneth A Jacobson, Jack Rose, and Rolf Zeller. Hematoxylin and Eosin Staining of Tissue and Cell Sections. *CSH Protocols*, 3(5):4986–4988, 2008.

[29] Famke Aeffner, Mark D Zarella, Nathan Buchbinder, Marilyn M Bui, Matthew R Goodman, Douglas J Hartman, Giovanni M Lujan, Mariam A Molani, Anil V Parwani, Kate Lillard, Oliver C Turner, Venkata N P Vemuri, Ana G Yuil-Valdes, and Douglas Bowman. Introduction to Digital Image Analysis in Whole-slide Imaging: A White Paper from the Digital Pathology Association. *Journal of pathology informatics*, 10:9, mar 2019.

[30] Kaustav Bera, Kurt A. Schalper, David L. Rimm, Vamsidhar Velcheti, and Anant Madabhushi. Artificial intelligence in digital pathology? New tools for diagnosis and precision oncology. *Nature Reviews Clinical Oncology*, 16(11):703–715, 2019.

[31] Liron Pantanowitz, Navid Farahani, and Anil V Parwani. Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. *Pathology and Laboratory Medicine International*, 7:23–33, 2015.

[32] J. D. Pallua, A. Brunner, B. Zelger, M. Schirmer, and J. Haybaeck. The future of pathology is digital. *Pathology Research and Practice*, 216(9):153040, 2020.

[33] B. Widrow and M. A. Lehr. 30 years of adaptive neural networks: perceptron, madaline, and backpropagation. *Proceedings of the IEEE*, 78(9):1415–1442, 1990.

[34] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[35] F. Rosenblatt. Perceptron simulation experiments. *Proceedings of the IRE*, 48(3):301–309, 1960.

[36] Tianping Chen and Hong Chen. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Transactions on Neural Networks*, 6(4):911–917, 1995.

[37] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

[38] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[39] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[40] Y. . Zhou, R. Chellappa, A. Vaid, and B. K. Jenkins. Image restoration using a neural network. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(7):1141–1151, 1988.

[41] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(9):533–536, 1986.

[42] Yann le Cun. A theoretical framework for Back-Propagation. In *Proceedings of the 1988 Connectionist Models Summer School*, pages 21–28, 1988.

[43] Y LeCun, B Boser, J S Denker, D Henderson, R E Howard, W Hubbard, and L D Jackel. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551, dec 1989.

[44] Robert Heicht-Nielsen. III.3 - Theory of the Backpropagation Neural Network. In Harry B T Neural Networks for Perception Wechsler, editor, *Neural Networks for Perception, Computation, Learning, and Architectures*, pages 65–93. Academic Press, 1992.

[45] Léon Bottou. Large-Scale Machine Learning with Stochastic Gradient Descent. In Yves Lechevallier and Gilbert Saporta, editors, *19th International Conference on Computational Statistics*, pages 177–186, Heidelberg, 2010. Physica-Verlag HD.

[46] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *ICLR*, pages 1–15, 2015.

[47] Quoc V Le, Adam Coates, Bobby Prochnow, and Andrew Y Ng. On Optimization Methods for Deep Learning. In *Proceedings of the 28th International Con- ference on Machine Learning*, 2011.

[48] Adityanarayanan Radhakrishnan and Caroline Uhler. Memorization in Overparameterized Autoencoders. *arXiv*, pages 1–25, 2019.

[49] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv preprint*, (Ml):1–14, 2013.

[50] Carl Doersch. Tutorial on Variational Autoencoders. pages 1–23, 2016.

[51] Zhou Wang, A C Bovik, H R Sheikh, and E P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, apr 2004.

[52] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. ?-VAE: Learning basic visual concepts with a constrained variational framework. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, pages 1–22, 2017.

[53] Grace Lindsay. Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *Journal of Cognitive Neuroscience*, pages 1–15, 2020.

[54] Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. NeuroImage Seeing it all : Convolutional network layers map the function of the human visual system. *NeuroImage*, 152(January 2016):184–194, 2017.

[55] B Zhou, D Bau, A Oliva, and A Torralba. Interpreting Deep Visual Representations via Network Dissection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2131–2145, 2019.

[56] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding Neural Networks Through Deep Visualization. *arXiv*, 2015.

[57] Marlene R Cohen and Adam Kohn. Measuring and interpreting neuronal correlations. *Nature Neuroscience*, 14(7), 2011.

[58] Homayoun Nazeran, Feng Rice, and William Moran. Biomedical image processing in pathology: a review. *Australian Physical & Engineering Sciences*, (April), 1995.

[59] H E Dytch and G L Wied. Artificial neural networks and their use in quantitative pathology. *Analytical and quantitative cytology and histology*, 12(6):379–393, 1990.

[60] Ronald S. Weinstein. Prospects for telepathology. *Human Pathology*, 17(5):433–434, 1986.

[61] Peter Bankhead, Maurice B. Loughrey, José A. Fernández, Yvonne Dombrowski, Darragh G. McArt, Philip D. Dunne, Stephen McQuaid, Ronan T. Gray, Liam J. Murray, Helen G. Coleman, Jacqueline A. James, Manuel Salto-Tellez, and Peter W. Hamilton. QuPath: Open source software for digital pathology image analysis. *Scientific Reports*, 7(1):1–7, 2017.

[62] Josh Moore, Melissa Linkert, Colin Blackburn, Mark Carroll, Richard K Ferguson, Helen Flynn, Kenneth Gillen, Roger Leigh, Simon Li, Dominik Lindner, William J Moore, Andrew J Patterson, Blazej Pindelski, Balaji Ramalingam, Emil Rozbicki,

Aleksandra Tarkowska, Petr Walczysko, Chris Allan, Jean-Marie Burel, and Jason Swedlow. OMERO and Bio-Formats 5: flexible access to large bioimaging datasets at scale. In *Proc.SPIE*, volume 9413, mar 2015.

[63] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology (Poznan, Poland)*, 19(1A):A68–A77, 2015.

[64] David R.J. Snead, Yee Wah Tsang, Aisha Meskiri, Peter K. Kimani, Richard Crossman, Nasir M. Rajpoot, Elaine Blessing, Klaus Chen, Kishore Gopalakrishnan, Paul Matthews, Navid Momtahan, Sarah Read-Jones, Shatrughan Sah, Emma Simmons, Bidisa Sinha, Sari Suortamo, Yen Yeo, Hesham El Daly, and Ian A. Cree. Validation of digital pathology imaging for primary histopathological diagnosis. *Histopathology*, 68(7):1063–1072, 2016.

[65] Sanjay Mukhopadhyay, Michael D Feldman, Esther Abels, Raheela Ashfaq, Senda Beltaifa, Nicolas G Cacciabeve, Helen P Cathro, Liang Cheng, Kumarasen Cooper, Glenn E Dickey, Ryan M Gill, Robert P Heaton Jr, René Kerstens, Guy M Lindberg, Reenu K Malhotra, James W Mandell, Ellen D Manlucu, Anne M Mills, Stacey E Mills, Christopher A Moskaluk, Mischa Nelis, Deepa T Patil, Christopher G Przybycin, Jordan P Reynolds, Brian P Rubin, Mohammad H Saboorian, Mauricio Salicru, Mark A Samols, Charles D Sturgis, Kevin O Turner, Mark R Wick, Ji Y Yoon, Po Zhao, and Clive R Taylor. Whole Slide Imaging Versus Microscopy for Primary Diagnosis in Surgical Pathology: A Multicenter Blinded Randomized Noninferiority Study of 1992 Cases (Pivotal Study). *The American journal of surgical pathology*, 42(1):39–52, jan 2018.

[66] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(January), 2019.

[67] Richard Colling, Helen Pitman, Karin Oien, Nasir Rajpoot, Philip Macklin, Velicia Bachtiar, Richard Booth, Alyson Bryant, Joshua Bull, Jonathan Bury, Fiona Carragher, Richard Colling, Graeme Collins, Clare Craig, Maria Freitas da Silva, Daniel Gosling, Jaco Jacobs, Lena Kajland-Wilén, Johanna Karling, Darragh Lawler, Stephen Lee, Philip Macklin, Keith Miller, Guy Mozolowski, Richard Nicholson, Daniel O'Connor, Mikkel Rahbek, Nasir Rajpoot, Alan Sumner, Dirk Vossen, Kieron White, Charlotte Wing, Corrina Wright, David Snead, Tony Sackville, and Clare Verrill. Artificial intelligence in digital pathology: a roadmap to routine use in clinical practice. *Journal of Pathology*, 249(2):143–150, 2019.

[68] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[69] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C Nelson, Jessica L Mega, and Dale R

Webster. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. 94043, 2020.

[70] Morgan P. McBee, Omer A. Awan, Andrew T. Colucci, Comeron W. Ghobadi, Nadja Kadom, Akash Kansagra, Srini Tridandapani, and William F. Auffermann. Deep Learning in Radiology. *Academic Radiology*, 25(11):1472–1480, 2018.

[71] Muhammad Khalid Khan Niazi, Anil V. Parwani, and Metin N. Gurcan. Digital pathology and artificial intelligence. *The Lancet Oncology*, 20(5):e253–e261, 2019.

[72] Hye Yoon Chang, Chan Kwon Jung, Junwoo Isaac Woo, Sanghun Lee, Joonyoung Cho, Sun Woo Kim, and Tae-Yeong Kwak. Artificial Intelligence in Pathology. *Journal of pathology and translational medicine*, 53(1):1–12, jan 2019.

[73] Anant Madabhushi and George Lee. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis*, 33:170–175, 2016.

[74] Dayong Wang and Aditya Khosla. Deep Learning for Identifying Metastatic Breast Cancer. *arXiv*, pages 1–6, 2016.

[75] Y. Al-Kofahi, W. Lassoued, W. Lee, and B. Roysam. Improved automatic detection and segmentation of cell nuclei in histopathology images. *IEEE Transactions on Biomedical Engineering*, 57(4):841–852, 2010.

[76] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A.W.M. Van Der Laak, Meyke Hermsen, Quirine F. Manson, Maschenka Balkenhol, Oscar Geessink, Nikolaos Stathonikos, Marcory C.R.F. Van Dijk, Peter Bult, Francisco Beca, Andrew H. Beck, Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, Aoxiao Zhong, Qi Dou, Quanzheng Li, Hao Chen, Huang Jing Lin, Pheng Ann Heng, Christian Haß, Elia Bruni, Quincy Wong, Ugur Halici, Mustafa Ümit Öner, Rengul Cetin-Atalay, Matt Berseth, Vitali Khvatkov, Alexei Vylegzhanin, Oren Kraus, Muhammad Shaban, Nasir Rajpoot, Ruqayya Awan, Korsuk Sirinukunwattana, Talha Qaiser, Yee Wah Tsang, David Tellez, Jonas Annuscheit, Peter Hufnagl, Mira Valkonen, Kimmo Kartasalo, Leena Latonen, Pekka Ruusuvuori, Kaisa Liimatainen, Shadi Albarqouni, Bharti Mungal, Ami George, Stefanie Demirci, Nassir Navab, Seiryo Watanabe, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, Hady Ahmady Phoulady, Vassili Kovalev, Alexander Kalinovsky, Vitali Liauchuk, Gloria Bueno, M. Milagro Fernandez-Carrobles, Ismael Serrano, Oscar Deniz, Daniel Racoceanu, and Rui Venâncio. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA - Journal of the American Medical Association*, 318(22):2199–2210, 2017.

[77] Kun-hsing Yu, Ce Zhang, Gerald J Berry, Russ B Altman, Christopher Re, Daniel L Rubin, and Michael Snyder. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature Communications*, 7(12474):1–10, 2016.

[78] Mehmet Günhan Ertosun and Daniel L Rubin. Automated Grading of Gliomas using Deep Learning in Digital Pathology Images: A modular approach with ensemble of convolutional neural networks. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2015:1899–1908, nov 2015.

[79] Shallu and Rajesh Mehra. Breast cancer histology images classification: Training from scratch or transfer learning? *ICT Express*, 4(4):247–254, 2018.

[80] Pegah Khosravi, Ehsan Kazemi, Marcin Imielinski, Olivier Elemento, and Iman Hajirasouliha. Deep Convolutional Neural Networks Enable Discrimination of Heterogeneous Digital Pathology Images. *EBioMedicine*, 27:317–328, 2018.

[81] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L. Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine*, page 1, 2018.

[82] Andreas Holzinger, Bernd Malle, Peter Kieseberg, Peter M Roth, M Heimo, Reihs Robert, and Kurt Zatloukal. Towards the Augmented Pathologist : Challenges of Explainable-AI in Digital Pathology. *arXiv*, pages 1–34, 2017.

[83] Andrew Janowczyk and Anant Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of Pathology Informatics*, 7(1), 2016.

[84] Po Hsuan Cameron Chen, Krishna Gadepalli, Robert MacDonald, Yun Liu, Shiro Kadowaki, Kunal Nagpal, Timo Kohlberger, Jeffrey Dean, Greg S. Corrado, Jason D. Hipp, Craig H. Mermel, and Martin C. Stumpe. An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. *Nature Medicine*, 25(9):1453–1457, 2019.

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

[86] Marc Macenko, Marc Niethammer, J S Marron, David Borland, John T Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E Thomas. A method for normalizing histology slides for quantitative analysis. *ISBI*, pages 1107–1110, 2009.

[87] A Vahadane, T Peng, A Sethi, S Albarqouni, L Wang, M Baust, K Steiger, A M Schlitter, I Esposito, and N Navab. Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images. *IEEE Transactions on Medical Imaging*, 35(8):1962–1971, 2016.

[88] Marc Niethammer, David Borland, J S Marron, John Woosley, and Nancy E Thomas. Appearance Normalization of Histology Slides. In Fei Wang, Pingkun Yan, Kenji Suzuki, and Dinggang Shen, editors, *Machine Learning in Medical Imaging*, pages 58–66, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[89] A M Khan, N Rajpoot, D Treanor, and D Magee. A Nonlinear Mapping Approach to Stain Normalization in Digital Histopathology Images Using Image-Specific Color Deconvolution. *IEEE Transactions on Biomedical Engineering*, 61(6):1729–1738, 2014.

[90] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. *arXiv preprint*, pages 1–9, 2014.

[91] Yair Rivenson, Hongda Wang, Zhensong Wei, Kevin de Haan, Yibo Zhang, Yichen Wu, Harun Günayd?n, Jonathan E Zuckerman, Thomas Chong, Anthony E Sisk, Lindsey M Westbrook, W Dean Wallace, and Aydogan Ozcan. Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning. *Nature Biomedical Engineering*, 3(6):466–477, 2019.

[92] Erik A Burlingame, Adam Margolin, Joe W Gray, and Young Hwan Chang. SHIFT: speedy histopathological-to-immunofluorescent translation of whole slide images using conditional generative adversarial networks. In *SPIE*, volume 10581, pages 1058105–1058107, 2018.

[93] J H Lagendijk, H Mullink, P J Van Diest, G A Meijer, and C J L M Meijer. Immunohistochemical differentiation between primary adenocarcinomas of the ovary and ovarian metastases of colonic and breast origin . Comparison between a statistical and an intuitive approach. *Journal of Clinical Pathology*, 52:283–290, 1999.

[94] J A R Amos Ara. Technical Aspects of Immunohistochemistry. *Vet Pathol*, 426:405–426, 2005.

[95] Mehrdad Nadji, Seyed Z Tabei, Albert Castro, T Ming Chu, and Azorides R Morales. Prostatic Origin of Tumors: An Immunohistochemical Study. *American Journal of Clinical Pathology*, 73(6):735–739, jun 1980.

[96] David F Steiner, Robert Macdonald, Yun Liu, Peter Truszkowski, Jason D Hipp, Christopher Gammage, Florence Thng, Lily Peng, and Martin C Stumpe. Impact of Deep Learning Assistance on the Histopathologic Review of Lymph Nodes for Metastatic Breast Cancer. *Am J Surg Pathol*, 42(12):1636–1646, 2018.

[97] Joel Saltz, Rajarsi Gupta, Le Hou, Tahsin Kurc, Pankaj Singh, Vu Nguyen, Dimitris Samaras, Kenneth R. Shroyer, Tianhao Zhao, ReJoel Saltz, Rajarsi Gupta, Le Hou, Tahsin Kurc, Pankaj Singh, Vu Nguyen, Dimitris Samaras, Kenneth R. Shroyer, Tianhao Zhao, Rebecca Batiste, John Van Arnam, Ilya The Cancer Genome Atlas Research Network, Shmulevich, Arvind U.K. Rao, Alexander J. Lazar, Ashish Sharma, and Vesteinn Thorsson. Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images. *Cell Reports*, 23(1):181–193.e7, 2018.

[98] Hamid Reza Tizhoosh and Liron Pantanowitz. Artificial Intelligence and Digital Pathology: Challenges and Opportunities. *Journal of pathology informatics*, 9:38, nov 2018.

[99] Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-michael Agapow, Michael Zietz, Michael M Hoffman, Wei Xie, Gail L Rosen, Benjamin J Lengerich, Johnny Israeli, Jack Lanchantin, Stephen Woloszynek, Anne E Carpenter, Avanti Shrikumar, Jinbo Xu, Evan M Cofer, Christopher A Lavender, Srinivas C Turaga, Amr M Alexandari, Zhiyong Lu, David J Harris, Dave Decaprio, Yanjun Qi, Anshul Kundaje, Yifan Peng, Laura K Wiley, Austin Huang, Anthony Gitter, and Casey S Greene. Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society Interface*, 15, 2018.

[100] Fereshteh Farzianpour, Sara Amirian, and Raziye Byravan. An Investigation on the Barriers and Facilitators of the Implementation of Electronic Health Records (EHR). *Scientific Research Publishing*, (December):1665–1670, 2015.

[101] Andrea Lynne Barbieri, Oluwole Fadare, Linda Fan, Hardeep Singh, and Vinita Parkash. Challenges in communication from referring clinicians to pathologists in the electronic health record era. *Journal of Pathology Informatics*, 9(1):1–6, 2018.

[102] Scott Mayer Mckinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg C Corrado, Ara Darzi, Mozziyar Etemadi, Florencia Garcia-vicente, Fiona J Gilbert, Mark Halling-brown, Demis Hassabis, Sunny Jansen, and Alan Karthikesalingam. International evaluation of an AI system for breast cancer screening. *Nature*, 577(January), 2020.

[103] Anshul Kundaje, Casey S Greene, Michael M Hoffman, and Jeffrey T Leek. The importance of transparency and reproducibility in artificial intelligence research. *arXiv*, 2020.

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

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

[106] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *arXiv*, pages 249–259, 2018.

[107] David M Metter, Terence J Colgan, Stanley T Leung, Charles F Timmons, and Jason Y Park. Trends in the US and Canadian Pathologist Workforces From 2007 to 2017. *JAMA - Journal of the American Medical Association*, 2(5):1–11, 2020.

[108] Luke Ternes, Ge Huang, Christian Lanciault, Guillaume Thibault, Rachelle Riggers, Joe W Gray, John Muschler, and Young Hwan Chang. VISTA: Virtual ImmunoSTAining for pancreatic disease quantification in murine cohorts. *bioRxiv*, page 2020.04.01.020842, jan 2020.

[109] Geoffrey F Schau, Guillaume Thibault, Mark A Dane, Joe W Gray, Laura M Heiser, and Young Hwan Chang. Variational autoencoding tissue response to microenvironment perturbation. In *Proc.SPIE*, volume 10949, mar 2019.

[110] Chun Han Lin, Tiina Jokela, Joe Gray, and Mark A. LaBarge. Combinatorial Microenvironments Impose a Continuum of Cellular Responses to a Single Pathway-Targeted Anti-cancer Compound. *Cell Reports*, 21(2):533–545, 2017.

[111] Melissa R Junttila and Frederic J de Sauvage. Influence of tumour micro-environment heterogeneity on therapeutic response. *Nature*, 501:346, sep 2013.

[112] Gianluca Pegoraro and Tom Misteli. High-Throughput Imaging for the Discovery of Cellular Mechanisms of Disease. *Trends in Genetics*, 33(9):604–615, 2017.

[113] Juan C. Caicedo, Sam Cooper, Florian Heigwer, Scott Warchal, Peng Qiu, Csaba Molnar, Aliaksei S. Vasilevich, Joseph D. Barry, Harmanjit Singh Bansal, Oren Kraus, Mathias Wawer, Lassi Paavolainen, Markus D. Herrmann, Mohammad Rohban, Jane Hung, Holger Hennig, John Concannon, Ian Smith, Paul A. Clemons, Shantanu Singh, Paul Rees, Peter Horvath, Roger G. Linington, and Anne E. Carpenter. Data-analysis strategies for image-based cell profiling. *Nature Methods*, 14(9):849–863, 2017.

[114] Chun-Han Lin, Jonathan K. Lee, and Mark A. LaBarge. Fabrication and Use of MicroEnvironment microArrays (MEArrays). *Journal of Visualized Experiments*, (68):1–7, 2012.

[115] Spencer S. Watson, Mark Dane, Koei Chin, Zuzana Tatarova, Moqing Liu, Tiera Liby, Wallace Thompson, Rebecca Smith, Michel Nederlof, Elmar Bucher, David Kilburn, Matthew Whitman, Damir Sudar, Gordon B. Mills, Laura M. Heiser, Oliver Jonas, Joe W. Gray, and James E. Korkola. Microenvironment-Mediated Mechanisms of Resistance to HER2 Inhibitors Differ between HER2+ Breast Cancer Subtypes. *Cell Systems*, 6(3):329–342.e6, 2018.

[116] Peter Goldsborough, Nick Pawlowski, Juan C Caicedo, Shantanu Singh, and Anne Carpenter. CytoGAN: Generative Modeling of Cell Images. *bioRxiv*, (Nips):227645, 2017.

[117] Philipp Eulenberg, Niklas Köhler, Thomas Blasi, Andrew Filby, Anne E. Carpenter, Paul Rees, Fabian J. Theis, and F. Alexander Wolf. Reconstructing cell cycle and disease progression using deep learning. *Nature Communications*, 8(1):1–6, 2017.

[118] Eric M. Christiansen, Samuel J. Yang, D. Michael Ando, Ashkan Javaherian, Gaia Skibinski, Scott Lipnick, Elliot Mount, Alison O'Neil, Kevan Shah, Alicia K. Lee, Piyush Goyal, William Fedus, Ryan Poplin, Andre Esteva, Marc Berndl, Lee L. Rubin, Philip Nelson, and Steven Finkbeiner. In Silico Labeling: Predicting Fluorescent Labels in Unlabeled Images. *Cell*, 173(3):792–795.e19, 2018.

[119] Zhiting Hu, Zichao Yang, Ruslan Salakhutdinov, and Eric P Xing. On unifying deep generative models. *arXiv preprint arXiv:1706.00550*, 2017.

[120] François Chollet et al. Keras. `https://github.com/fchollet/keras`, 2015.

[121] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, Xiaoqiang Zheng, and Google Brain. TensorFlow: A System for Large-Scale Machine Learning TensorFlow: A system for large-scale machine learning. *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)*, pages 265–284, 2016.

[122] L J P Van Der Maaten and G E Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[123] Joseph L. Fleiss. The Equivalence of Weighted Kappa and the Interclass Correlation Coefficient as Measures of Reliability. *Education and Psychological Measurement*, 33:613–619, 1973.

[124] Joan Massagué. Tgfβ in cancer. *Cell*, 134(2):215–230, 2008.

[125] Geoffrey F Schau, Erik A Burlingame, Guillaume Thibault, Tauangtham Anekpuritanang, Ying Wang, Joe W Gray, Christopher Corless, and Young Hwan Chang. Predicting primary site of secondary liver cancer with a neural estimator of metastatic origin. *Journal of Medical Imaging*, 7(1):1–9, feb 2020.

[126] Ashwin Ananthakrishnan, Veena Gogineni, Kia Saeian, and M Sc. Epidemiology of Primary and Secondary Liver Cancers. *Seminars in Interventional Radiology*, 23(1):47–63, 2006.

[127] M Mohammadian, N Mahdavifar, and H Salehiniya. Liver Cancer in the World: Epidemiology, Incidence, Mortality and Risk Factors. *World Cancer Research Journal*, 5(2), 2018.

[128] Ugljesa Djuric, Gelareh Zadeh, Kenneth Aldape, and Phedias Diamandis. Precision histology : how deep learning is poised to revitalize histomorphology for personalized cancer care. *npj Precision Oncology*, (April):1–4, 2017.

[129] Geert Litjens, Clara I Sánchez, Nadya Timofeeva, Meyke Hermsen, Iris Nagtegaal, Iringo Kovacs, Christina Hulsbergen Van De Kaa, Peter Bult, and Bram Van Ginneken. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Nature Publishing Group*, (January):1–11, 2016.

[130] Jakob Nikolas Kather. Predicting survival from colorectal cancer histology slides using deep learning : A retrospective multicenter study. *Plos Medicine*, pages 1–14, 2019.

[131] Angel Cruz-roa, Hannah Gilmore, Ajay Basavanhally, Michael Feldman, Shridar Ganesan, Natalie N C Shih, John Tomaszewski, and Fabio A González. Accurate and reproducible invasive breast cancer detection in whole- slide images : A Deep Learning approach for quantifying tumor extent. *Nature Publishing Group*, (April):1–14, 2017.

[132] Jeffrey J Nirschl, Andrew Janowczyk, Eliot G Peyster, Renee Frank, B Margulies, Michael D Feldman, and Anant Madabhushi. A deep-learning classifier identifies patients with clinical heart failure using whole-slide images of H & E tissue. *PLoS ONE*, pages 1–16, 2018.

[133] Bruno Korbar, Andrea M Olofson, Allen P Miraflor, Catherine M Nicka, and Matthew A Suriawinata. Deep Learning for Classification of Colorectal Polyps on Whole-slide Images. *J Pathol Inform.*, 8(30), 2017.

[134] Zahraa Al-milaji, Ilker Ersoy, Adel Hafiane, Kannappan Palaniappan, and Filiz Bunyak. Integrating segmentation with deep learning for enhanced classification of epithelial and stromal tissues in H & E images. *Pattern Recognition Letters*, 119:214–221, 2019.

[135] Yushan Zheng, Zhiguo Jiang, Fengying Xie, Haopeng Zhang, and Yibing Ma. Feature extraction from histopathological images based on nucleus-guided convolutional neural network for breast lesion classification. *Pattern Recognition*, 71:14–25, 2017.

[136] A. Garcia-Garcia, S. Orts-Escolano, S.O. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez. A Review on Deep Learning Techniques Applied to Semantic Segmentation. *arXiv*, pages 1–23.

[137] Christof Angermueller, Tanel Pärnamaa, Leopold Parts, and Oliver Stegle. Deep learning for computational biology. *Molecular Systems Biology*, 12(878):1–16, 2016.

[138] Dina Aboul Dahab, Samy S A Ghoniemy, and Gamal M Selim. Automated Brain Tumor Detection and Identification Using Image Processing and Probabilistic Neural Network Techniques. *International Journal of Image Processing and Visual Communication*, 1(2):1–8, 2012.

[139] Mohammad Havaei, Axel Davy, David Warde-farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with Deep Neural Networks. *Medical Image Analysis*, 35:18–31, 2017.

[140] Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-jones, A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Wei Xie, Gail L Rosen, Benjamin J Lengerich, and Johnny Israeli. Opportunities and obstacles for deep learning in biology and medicine. *bioRxiv*, 2017.

[141] Muhammad Khalid, Khan Niazi, Anil V Parwani, and Metin N Gurcan. Series Digital Oncology 1 Digital pathology and artificial intelligence. *Lancet Oncology*, 20(5):e253–e261, 2019.

[142] Chris Allan, Jean-Marie Burel, Josh Moore, Colin Blackburn, Melissa Linkert, Scott Loynton, Donald MacDonald, William J Moore, Carlos Neves, Andrew Patterson, Michael Porter, Aleksandra Tarkowska, Brian Loranger, Jerome Avondo, Ingvar Lagerstedt, Luca Lianas, Simone Leo, Katherine Hands, Ron T Hay, Ardan Patwardhan, Christoph Best, Gerard J Kleywegt, Gianluigi Zanetti, and Jason R Swedlow.

OMERO: flexible, model-driven data management for experimental biology. *Nature Methods*, 9:245, feb 2012.

[143] Adam Goode, Benjamin Gilbert, Jan Harkes, Drazen Jukic, and Mahadev Satyanarayanan. OpenSlide: A vendor-neutral software foundation for digital pathology. *Journal of Pathology Informatics*, 4(1), 2013.

[144] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *arXiv*, 2016.

[145] Leslie N Smith. Cyclical Learning Rates for Training Neural Networks. *arXiv*, (April), 2015.

[146] E. Meijering. Cell segmentation: 50 years down the road [life sciences]. *IEEE Signal Processing Magazine*, 29(5):140–145, 2012.

[147] Anne E. Carpenter, Thouis R. Jones, Michael R. Lamprecht, Colin Clarke, In Han Kang, Ola Friman, David A. Guertin, Joo Han Chang, Robert A Lindquist, Jason Moffat, Polina Golland, and David M. Sabatini. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol*, 7(10):R100, 2006.

[148] C Sommer, C Straehle, U Köthe, and F A Hamprecht. Ilastik: Interactive learning and segmentation toolkit. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 230–233, 2011.

[149] Li Fei-Fei, R Fergus, and P Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.

[150] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *34th International Conference on Machine Learning*, 2017.

[151] Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic Model-Agnostic Meta-Learning. In *32nd Conference on Neural Information Processing Systems*, number NeurIPS, 2018.

[152] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical Networks for Few-shot Learning. In *31st Conference on Neural Information Processing Systems*, 2017.

[153] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[154] Brenden M Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua B Tenenbaum. One shot learning of simple visual concepts. In *Proceedings of the Annual Meeting of the Cognitive Science*, 2011.

[155] Leland McInnes and John Healy. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*, pages 1–18, 2018.

[156] Meenhard Herlyn, Gloria Balaban, Jeannette Bennicelli, DuPont Guerry IV, Ruth Halaban, Dorothee Herlyn, David E Elder, Gerd G Maul, Zenon Steplewski, Peter C Nowell, Wallace H Clark, and Hilary Koprowski. Primary Melanoma Cells of the Vertical Growth Phase: Similarities to Metastatic Cells2. *JNCI: Journal of the National Cancer Institute*, 74(2):283–289, feb 1985.

[157] Man-Hung Eric Tang, Malin Dahlgren, Christian Brueffer, Tamara Tjitrowirjo, Christof Winter, Yilun Chen, Eleonor Olsson, Kun Wang, Therese Törngren, Martin Sjöström, Dorthe Grabau, Pär-Ola Bendahl, Lisa Rydén, Emma Niméus, Lao H Saal, Åke Borg, and Sofia K Gruvberger-Saal. Remarkable similarities of chromosomal rearrangements between primary human breast cancers and matched distant metastases as revealed by whole-genome sequencing. *Oncotarget*, 6(35):37169–37184, nov 2015.

[158] Wenyuan Dai, Giang Yang, Gui-Rong Xue, and Yong Yu. Boosting for Transfer Learning. In *Proceedings of the 24th International Conference on Machine Learning*, pages 193–200, 1997.

[159] Yoshua Bengio. Deep Learning of Representations for Unsupervised and Transfer Learning. In *JMLR: Workshop and Conference Proceedings 27*, pages 17–37, 2012.

[160] Yuqing Gao and Khalid M Mosalam. Deep Transfer Learning for Image-Based Structural Damage Recognition. *Computer-Aided Civil and Infrastructure Engineering*, 33(9):748–768, sep 2018.