Out-of-Distribution Generalization of Gigapixel Image Representation

by

Milad Sikaroudi

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2023

© Milad Sikaroudi 2023

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Member:	Majid Ahmadi Prof., Dept. Electrical and Computer Engineering, University of Windsor
Internal-External Members:	Mohammad Kohandel Prof., Dept. Applied Mathematics, University of Waterloo
	Mark Crowley Associate Prof., Dept. Electrical and Computer Engineering, University of Waterloo
Internal Member:	Jonathan Kofman Prof., Dept. Systems Design Engineering, University of Waterloo
Supervisor(s):	Hamid Tizhoosh Prof., Mayo Clinic
	Shahryar Rahnamayan Prof., Brock University

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Every piece of experimental data, graphical representations, and textual content presented here is the product of my own efforts during my Ph.D. research journey. These constitute the collective body of contributions I made throughout my doctorate studies. Whether or not the thesis directly draws upon the following papers that I have authored is a variable factor:

a) **Sikaroudi M,** Hosseini M, Rahnamayan S, Tizhoosh HR. ALFA–Leveraging All Levels of Feature Abstraction for Enhancing the Generalization of Histopathology Image Classification Across Unseen Hospitals. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2023 Aug.

b) **Sikaroudi M,** Hosseini M, Gonzalez R, Rahnamayan S, Tizhoosh HR. Generalization of Vision Pre-trained Models for Histopathology. Nature, Scientific Reports. 2023 Apr.

c) Alsaafin A, Safarpoor A, **Sikaroudi M**, Hipp JD, Tizhoosh HR. Learning to Predict RNA Sequence Expressions from Whole Slide Images with Applications for Search and Classification. Nature, Communications Biology. 2023 Mar.

d) Hosseini SM, **Sikaroudi M**, Babaie M, Tizhoosh HR. Proportionally Fair Hospital Collaborations in Federated Learning of Histopathology Images. IEEE Transactions on Medical Imaging. 2023 Jan.

e) **Sikaroudi M**, Rahnamayan S, Tizhoosh HR. Hospital-agnostic image representation learning in digital pathology. 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) 2022 Jul.

f) Hosseini SM, **Sikaroudi M**, Babaei M, Tizhoosh HR. Cluster-Based Secure Multiparty Computation in Federated Learning for Histopathology Images. MICCAI Workshop, International Workshop on Distributed, Collaborative, and Federated Learning 2022 Sep.

g) **Sikaroudi M**, Ghojogh B, Karray F, Crowley M, Tizhoosh HR. Magnification Generalization for Histopathology Image Embedding. 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) 2021 Apr.

h) **Sikaroudi M**, Ghojogh B, Karray F, Crowley M, Tizhoosh HR. Batch-incremental Triplet Sampling for Training Triplet Networks using Bayesian Updating Theorem. 2020 25th IEEE International Conference on Pattern Recognition (ICPR) 2021 Jan.

i) **Sikaroudi M**, Ghojogh B, Safarpoor A, Karray F, Crowley M, Tizhoosh HR. Offline versus Online Triplet Mining based on Extreme Distances of Histopathology Patches. International Symposium on Visual Computing 2020 Oct.

j) **Sikaroudi M**, Safarpoor A, Ghojogh B, Shafiei S, Crowley M, Tizhoosh HR. Supervision and Source Domain Impact on Representation Learning: A Histopathology Case Study. 2020 42nd IEEE Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) 2020 Jul.

k) Ghojogh B, **Sikaroudi M**, Shafiei S, Tizhoosh HR, Karray F, Crowley M. Fisher Discriminant Triplet and Contrastive Losses for Training Siamese Networks. 2020 IEEE International Joint Conference on Neural Networks (IJCNN) 2020 Jul.

l) Ghojogh B, **Sikaroudi M**, Tizhoosh HR, Karray F, Crowley M. Weighted Fisher Discriminant Analysis in the Input and Feature Spaces. International Conference on Image Analysis and Recognition 2020 Jun.

Abstract

This thesis addresses the significant challenge of improving the generalization capabilities of artificial deep neural networks in the classification of whole slide images (WSIs) in histopathology across different and unseen hospitals. It is a critical issue in AI applications to vision-based healthcare tasks, given that current standard methodologies struggle with out-of-distribution (OOD) data from varying hospital sources. In histopathology, distribution shifts can arise due to image acquisition variances across different scanner vendors, differences in laboratory routines and staining procedures, and diversity in patient demographics. This work investigates two critical forms of generalization within histopathology: magnification generalization and OOD generalization towards different hospitals. One chapter of this thesis is dedicated to the exploration of magnification generalization, acknowledging the variability in histopathological images due to distinct magnification levels and seeking to enhance the model's robustness by learning invariant features across these levels. However, the major part of this work focuses on OOD generalization, specifically unseen hospital data. The objective is to leverage knowledge encapsulated in pre-existing models to help new models adapt to diverse data scenarios and ensure their efficient operation in different hospital environments. Additionally, the concept of Hospital-Agnostic (HA) learning regimes is introduced, focusing on invariant characteristics across hospitals and aiming to establish a learning model that sustains stable performance in varied hospital settings. The culmination of this research introduces a comprehensive method, termed ALFA (*Exploiting All Levels of Feature Abstraction*), that not only considers invariant features across hospitals but also extracts a broader set of features from input images, thus maximizing the model's generalization potential. The findings of this research are expected to have significant implications for the deployment of medical image classification systems using deep models in clinical settings. The proposed methods allow for more accurate and reliable diagnostic support across various hospital environments, thereby improving diagnostic accuracy and reliability, and paving the way for enhanced generalization in histopathology diagnostics using deep learning techniques. Future research directions may build on expanding these investigations to further improve generalization in histopathology.

Acknowledgements

Navigating the path of this thesis, I have been graced with guidance, support, and a sense of community from numerous quarters.

Firstly, I extend profound appreciation to my supervisors, Professor H.R. Tizhoosh and Prof. Shahryar Rahnamayan. Their steadfast support, sage advice, and patience have been foundational to my journey. Collaborating under their tutelage has been not just an honor but also a transformative experience.

My heartfelt thanks to the vibrant community at KIMIA Lab, with a special nod to Mitra whose unwavering dedication stood out. To my friends — Benyamin, Amir, Danial, Sobhan H., Maryam, Parsa, Sobhan S., and Morteza — the unity, cherished moments, and mutual challenges we've shared have etched a special place in my heart.

I'm deeply indebted to my committee members: Prof. Kohandel, Prof. Crowley, Prof. Kofman, and Prof. Ahmadi. Their insightful critiques, valuable guidance, and academic arsenal were essential to my progress.

A shoutout to my social circle, notably Hossein S., Hamid, Kasra, and Fatemeh, with whom I found solace and relaxation in between the rigors of research.

I send my deepest gratitude to friends from afar — Alireza, Hossein, Narges S., and Javad. Even separated by vast distances, their unwavering support and heartfelt prayers resonated with me. Their words of encouragement and periodic respites were like a balm, making every step of this journey smoother.

With emotions running deep and a heart full of warmth, I extend my profound thanks to my parents, my recently departed grandmother who remains a beacon in my heart, my treasured sisters, Sanaz and Solmaz, my dearest nieces, Arsham and Armin, and my unwavering brothers-in-law, Saeed and Mohammad. They've collectively embroidered a rich pattern of values in the fabric of my being — encompassing diligence, patience, dignity in upholding promises, and the timeless treasure of knowledge. Their steadfast trust and insights have been the guiding constellations in my explorative journey.

To my cherished Sanaz (Minooz), the rhythm of my heartbeat. Your unwavering support, the zest you bring, the solace in your care, and your enduring faith in me — even in moments I faltered — have been nothing short of poetic. Your presence in my life has elevated every experience. This milestone is a reflection of us, and I'm blessed to have you beside me, sharing in this accomplishment.

In conclusion, I pay my respects to the divine forces that endowed me with the resilience, grit, and ceaseless inspiration to overcome every hurdle.

Dedication

This is dedicated to my parents.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AUROC Area Under the ROC curve 78, 82

ccRCC Clear Cell Renal Cell Carcinoma 66, 69

CNN Convolutional Neural Network xiii, 2, 10, 11, 13–15, 22, 27, 62

COMPAS Correctional Offender Management Profiling for Alternative Sanctions 39

crRCC Chromophobe Renal Cell Carcinoma 66, 69

Cycle-GAN Cycle-Generative Adversarial Networks 57

DA Domain Adaptation 8, 18, 19, 88

DG Domain Generalization 6, 8, 18–21, 26, 27, 30, 37, 60, 62, 70, 72, 73, 79, 80, 82, 85, 88

DL Deep Learning 27

DML Deep Metric Learning 13, 21

DNN Deep Neural Network 8, 13, 15–17, 27, 70

ERM Expectation Risk Minimization xv, xvii, 26, 35–37, 42, 56, 59, 60, 67–71, 78–84

H&E Hematoxylin and Eosin 1, 9, 10

HED Hematoxylin and Eosin Dye xiv, 46, 75

i.i.d. Independent and Identically Distributed 2, 15, 17, 28, 39

IGC International Genomics Consortium Cancer Center xv, 66, 68, 70, 79, 81, 82, 84

KL Kullback–Leibler 22, 23, 32, 33, 70

MAML Model-Agnostic Meta-Learning 20–22, 24, 30, 36, 86

MASF Model-Agnostic Semantic learning of Features xiii, xvii, 20–22, 24, 30, 31, 36, 62, 64, 70

MDL Multi-Domain Learning 6, 8, 17, 18, 20, 27, 62

mDSDI meta Domain-Specific Domain-Invariant xv, xvi, 24, 25, 72, 73, 78–84

MIL Multiple-Instance Learning 86, 87

ML Machine Learning 5, 8, 10, 14, 16, 17, 21, 28, 61, 86–88

MRI Magnetic Resonance Imaging 15

MSKCC Memorial Sloan Kettering Cancer Center xv, 56, 66, 68, 70, 79, 82, 84

NCI National Cancer Institute Urologic Oncology Branch xv, 56, 66, 67, 70, 79, 81, 82, 84

OOD Out-Of-Distribution xiii, xiv, xvii, 2, 4–8, 16, 17, 26, 27, 29, 37–45, 48–55, 57–61, 85

PAD Proxy-A Distance 87

PCA Principal Component Analysis 69, 79

pRCC Papillary Renal Cell Carcinoma 66, 69

RCC Renal Cell Carcinoma xvii, 15, 65–67, 69–71, 73, 78–80, 82, 84

Semi-SL Semi Supervised Learning xvii, 27, 43, 47, 49, 51, 56, 58, 59

Semi-WSL Semi-Weakly-Supervised Learning xvii, 27, 43, 47, 49, 51, 56, 58, 59

SGD Stochastic Gradient Descent 22, 47

SSL Self Supervised Learning xv, 72–76, 80, 81, 83

TCGA The Cancer Genome Atlas xiii, 10, 15, 27, 29, 43, 44, 56, 59, 66, 73, 79, 82

UMAP Uniform Manifold Approximation and Projection 69, 79

ViT Vision Transformer 13

WSI Whole Slide Image xiii, xv, 1, 3, 5, 8–10, 16, 28–30, 41–44, 62, 63, 66, 67, 69, 70, 73, 81, 86, 87

XAI Explainable Artificial Intelligence 41, 51, 56, 58–60, 88

Chapter 1

Motivation

1.1 Cancer Diagnosis

Cancer is the world's major cause of death, estimated to claim almost 10 million lives in 2020 [164]. The early and reliable diagnosis of cancer is of high importance since it is conducive to a higher chance of successful treatment. Although a potential cancer diagnosis may be initiated based on medical history and physical examination, ultimate verification and final diagnosis need a biopsy and histopathologic inspection. Hematoxylin and Eosin (H&E) stained slides of biopsy samples on glass slides are routinely used by pathologists to examine human tissue under the microscope [37]. Histopathological visual inspections provide diagnostic and predictive information regarding disease progression and phenotypic characteristics. In most cases, pathologists may stratify tumors based on simple decision trees that they develop during their training. Inter-observer variability and prolonged diagnosis times are the two major limitations of examining H&E slides by pathologists [77]. These limitations can be overcome using computer-aided approaches, which can reveal nuances in morphology between clinical groups.

1.2 Digital Pathology

The advent of WSI scanners and the impact of the COVID-19 pandemic has started a revolution in diagnostic pathology [168]. The WSI scanner is capable of digitizing conventional glass tissue slides into digital images, significantly facilitating the application of image analysis in pathology. Image analysis, which uses computational techniques to interpret pathology images, is rapidly becoming a powerful tool for examining a wide variety of pathology workflows. Numerous studies [102, 231] have shown that such technologies can overcome the inherent subjectivity involved in manual analysis and significantly reduce pathologists' burden via high-throughput computerized analysis.

1.3 Deep Learning and Computational Pathology

Deep learning, particularly a CNN, has significantly improved the accuracy of many tasks such as image recognition, object detection, and semantic segmentation [158, 265, 3]. One of the primary tasks in computer vision is *image classification*. Deep learning approaches have demonstrated their effectiveness in this area, but their performance is ultimately constrained by the size of the available dataset [105]. The availability of large datasets has contributed to significant progress in image classification, facilitated by the emergence of deep learning techniques. Furthermore, CNNs have proven successful in capturing intricate tissue patterns and are extensively utilized in biomedical imaging for cancer detection and segmentation of breast, lung, and prostate cancers [105].

However, the limited availability of expert-labeled training data [231] and the challenge of generalizing beyond the training data [185] raise concerns about the applicability of these models in computational pathology.

Transfer learning is utilized to overcome this limitation. It is all about re-purposing previously learned abstract information in a new context, similar to humans in the way that they do not learn everything from scratch and instead use and transfer their knowledge from previously learned areas to new domains and tasks. In histopathology, the recurring histomorphologic patterns are a hallmark of organizing diseases into meaningful subgroups by pathologists. These shared histomorphologic patterns appearing in different types of cancers would help pathologists to approximate clinical conditions for varied types of diseases which would be outside of their subspecialty. But, the problem is the assumption that the source and target domains are Independent and Identically Distributed (i.i.d.) while disregarding OOD scenarios that occur frequently in reality. In most scenarios, even the smallest difference in the statistics of external validation data compared to the training data can cause a method to utterly fail [233].

1.4 Generalization in Digital Pathology

Histopathology images inherently possess certain characteristics that make them challenging to handle; they are very large, diverse in terms of subtypes and classes, and display inconsistencies and fluctuations due to varying standards in acquisition and processing. This leads to several generalization issues that need to be tackled:

Generalization to different magnification levels– This involves adapting our methods to accommodate varying scales of image magnification, i.e., the field of view and resolution to perceive tissue characteristics.

Generalization to unseen subtypes– This area of study falls under the *zero/few-shot learning* scheme, where the objective is to adapt to classes that were not visible during the training phase.

Generalization to the same classes but from different distributions– This is also known as *Out-Of-Distribution generalization*. It is a significant aspect of research in the field, particularly when considering the problem of generalizing to unseen hospitals, where the data might be derived from a different source set.

1.4.1 Generalization to Different Magnification Levels

Magnification generalization is essential in the analysis of histopathological WSI for several reasons:

Detailed examination– In histopathology, a detailed examination of tissue samples at different magnification levels can reveal more about the disease's extent and nature. For example, the subtleties of some cellular or tissue abnormalities might only be visible at higher magnifications.

Consistency across different settings– Magnification generalization allows for a more consistent interpretation of images across different settings and devices. Different laboratories might use different magnification levels based on their protocols and the equipment they use. Hence, a model that can be generalized across different magnifications can be more universally applicable.

Invariance to scale– Tissue samples can exhibit considerable variation in their appearance depending on the magnification level used. A model that can handle these variations (i.e., that is invariant to magnification level) is likely to be more robust and hence more accurate in tissue categorization.

Handling scarce labeled data– In the context of machine learning, models that can generalize across different magnification levels can help address the problem of scarce labeled data. If a model can learn to recognize patterns at different magnifications, it can effectively increase the amount of useful training data.

1.4.2 Generalization to Unseen Classes

Generalization to unseen classes which is being studied through the few-shot learning framework is vital for histopathology for several reasons:

Diversity of Disease Manifestations– Histopathology involves the examination of a wide range of diseases, each with unique cellular and tissue characteristics. These diseases may manifest differently in different patients and might not be adequately (or at all) represented in the training set. Therefore, the ability to generalize to unseen classes is crucial for accurate diagnosis and prognosis.

Data Scarcity– Well-annotated histopathological data is often scarce due to the time-consuming and expensive nature of data collection and annotation. This scarcity makes it difficult to train robust models using traditional supervised learning methods, which require

a large amount of labeled data. Few-shot learning methods, which are designed to learn from limited labeled data, are therefore essential for developing effective histopathology image classification systems.

Domain Shifts and Variations– Histopathological images can vary significantly due to factors such as different acquisition protocols, primary sites (organs), and tissue types. These variations pose significant challenges for traditional machine-learning methods. A model that can generalize well to unseen classes can handle these domain shifts and variations more effectively.

Robustness to Different Magnifications and Stains– Different laboratories may use different slide preparation protocols, staining techniques, and magnification levels when creating histopathological images. A model that can generalize to unseen classes and use few-shot learning can handle these variations and provide more consistent and reliable results.

Efficient Use of Available Data– Few-shot learning methods are designed to learn from limited data, making the most of the available labeled samples. This efficiency is particularly important in histopathology, where the collection and annotation of data can be extremely labor-intensive and costly due to the gigapixel nature of digitized tissue samples.

Adapting to Novel Diseases– Diseases can evolve over time, and new diseases can emerge. A model that can generalize to unseen classes and utilize few-shot learning can adapt more quickly to these novel diseases, helping to ensure timely and accurate diagnoses.

1.4.3 Generalization to OOD Data Coming from Different Hospitals

Generalization to unseen hospital data is paramount in histopathology due to the following reasons:

Variability in Data– Histopathology data varies significantly across different hospitals, laboratories, and regions due to differences in tissue preparation, staining protocols, imaging equipment, individual pathologist interpretation, and patient demographics. These variations can cause a model trained on data from one distribution (e.g., a specific hospital) to perform poorly when tested on data from a different distribution (e.g., a different hospital or lab). Thus, addressing OOD generalization is crucial for developing robust models that can perform well across diverse clinical settings.

Biases in Data– Histopathological data can often be biased due to factors such as differences in the patient population, the prevalence of certain diseases, and other demographic and socio-economic factors, the so-called social determinants. For instance, different races or sexes under different circumstances might have different histomorphologic patterns. If a model is trained on biased data, it may not perform well on data from an unseen distribution that does not share the same biases.

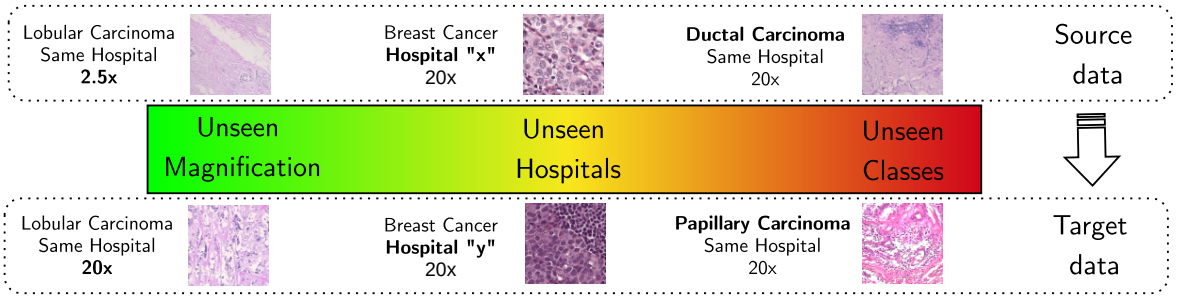


Figure 1.1: Different types of generalization applicable to digital pathology WSIs.

Scarcity of Labeled Data– Histopathology datasets often suffer from a scarcity of labeled data, which makes it challenging to properly train deep learning models that can generalize well to unseen data distributions. By focusing on OOD generalization, techniques like self-supervised learning and domain adaptation can be employed to learn useful representations from unlabeled or semi-labeled data, thereby improving the model's performance on unseen data.

Clinical Applicability– In clinical settings, a model must perform well on unseen data during training. This includes not just data from different hospitals or labs, but also data from new patients with potentially different disease manifestations or subtypes. Therefore, addressing OOD generalization is indispensable in ensuring that the models are clinically applicable and can provide reliable and accurate results across different scenarios.

By addressing these challenges, researchers can build more robust and reliable models for histopathology that can help in accurate disease diagnosis and prognosis, ultimately leading to better patient care.

1.5 Importance of Generalization in Histopathology

Generalization is a cornerstone of effective Machine Learning (ML) models and carries special significance in the field of histopathology. This is due to the essential role histopathological analysis plays in diagnosis, a critical process directly influencing patient outcomes, irrespective of factors such as race, ethnicity, or geographical location. To ensure patient safety and enhance the efficacy of treatment plans, it is crucial to develop robust models capable of handling the inherent distribution shift in histopathology data, thus exhibiting a "hospital-agnostic" behavior.

The types of generalization - magnification generalization, generalization to unseen classes, and OOD generalization - each cover vital aspects of this challenge in histopathology.

Magnification generalization, by allowing models to discern critical disease characteristics across different magnification levels, aids in the detailed and accurate examination

of tissue samples. This, in turn, enhances consistency in various laboratory settings and improves the robustness of deep models.

Generalization to unseen classes addresses the diversity of disease manifestations, the scarcity of annotated histopathological data, and the variations in image acquisition protocols. With the application of few-shot learning methods, models can effectively use available data, handle domain shifts, and adapt quickly to novel diseases, significantly contributing to timely and accurate diagnoses.

Finally, OOD generalization targets the crucial issue of data variability between different hospitals, laboratories, and regions. By focusing on this form of generalization, we can tackle potential biases in data, make the best use of scarce labeled data through techniques like self-supervised learning and domain adaptation, and enhance the clinical applicability of models.

While attempts to increase the diversity of training data and applications of stain normalization methods [178] can somewhat mitigate the distribution shift issue, this study will explore the potential of Multi-Domain Learning (MDL) regimes, particularly Domain Generalization (DG) techniques, in addressing these challenges.

In this thesis, the primary focus is on two essential types of generalization: magnification generalization and OOD generalization to unseen hospitals. The aim is to propose effective solutions and applications to address these specific challenges in histopathology.

The pursuit of magnification generalization revolves around building models that can effectively adapt to and interpret histopathological images at various magnification levels. Such models not only enhance the detail and precision of disease diagnoses but also promote consistency across different laboratory protocols and settings, contributing to their broader applicability.

In dealing with OOD generalization, the goal is to devise models capable of maintaining robust performance even when confronted with data from unseen hospitals. This aspect is of particular significance given the variations in data collection, preparation, and interpretation practices across different hospitals, laboratories, and regions.

By introducing and addressing these forms of generalization, the intention is to create machine learning models that exhibit remarkable adaptability to a wide array of conditions, efficiently manage variations in data, and offer consistent, reliable performance. Through this endeavor, the thesis strives to contribute to the overarching objective of improving diagnostic accuracy, enhancing prognosis assessment, and ultimately, elevating the quality of patient care in histopathology.

1.6 Outline of the Thesis

In this thesis, a structured approach is followed, addressing two fundamental types of generalization pertinent to the field of histopathology.

The relevant studies and methodologies related to this thesis are provided as background information in Chapter 2.

The first area of exploration centers on the challenge of magnification generalization. Accordingly, in Chapter 3 a method is proposed for emphasizing the learning of invariant features across different magnification levels. Recognizing the variability in histopathological images owing to distinct magnification levels, the proposed method aims to identify and learn those features that remain consistent, thus enhancing model robustness and reliability across varied settings and equipment.

In Chapter 4, the thesis attention is subsequently directed to OOD generalization, especially as it relates to unseen hospitals. This problem is formulated, and the potential contribution of pre-trained models in advancing OOD generalization is assessed. The aim is to use the knowledge housed within these pre-existing models to develop models that are effectively equipped to navigate and adapt to the diverse data scenarios encountered in different hospital environments.

In the subsequent phase of the investigation in Chapter 5, the thesis proposes a hospital-agnostic learning regime. This approach pivots on the idea of invariant features serving as the semantic core of the data across different hospitals. By emphasizing these consistent features, this research aspires to devise a learning model that maintains its performance stability across diverse hospital environments, ensuring its utility and reliability in varying clinical settings.

Finally, the thesis reaches its pinnacle with the proposal of a comprehensive method in Chapter 6. This method goes beyond focusing merely on invariant features across hospitals. It also considers an expanded set of features extractable from the input images, aiming to exploit all valuable information to maximize generalization potential. The thesis posits that this multi-faceted approach will significantly enhance the generalization capability of the model, delivering superior performance and contributing to the progress of diagnostic accuracy in histopathology.

Chapter 2

Background

In this chapter presents a review of the pertinent literature, aiming to shed light on the diverse range of topics relevant to our research. Our journey traverses through various interconnected domains, starting with digital histopathology and WSI, and ending with the exploration of methods that emphasize invariant and domain-specific features.

We embark on our exploration by delving into the realm of digital histopathology and WSI. In this section, we assess their fundamental role in transforming modern pathology and enhancing cancer diagnosis capabilities. We also review the evolution of these tools and discuss their growing influence on the application and development of ML tools in histopathology.

As our exploration progresses, we focus on the core concepts of representation and deep learning. We delve into how these tools have catalyzed advancements in cancer diagnosis and have become integral components of the ML toolkit. Their profound impact on the field necessitates an in-depth discussion to appreciate the depth of their influence.

Despite the significant strides made by Deep Neural Network (DNN) in disease diagnosis, they are not devoid of challenges. In the ensuing section, the thesis highlights these challenges and describes potential solutions. Among these, transfer learning emerges as a promising approach, capable of enhancing the robustness and performance of DNNs.

The journey through the literature next guides us towards a deep dive into transfer learning to critically assess the related studies. Along the way, the concept of generalization is introduced, a fundamental component of my research. The significance of this problem and its relevance in the context of machine learning applications in histopathology is explained.

In the following sections, the review scrutinizes MDL techniques, including Domain Adaptation (DA) and DG. The objective here is to present an exhaustive understanding of these techniques and their contributions toward addressing the generalization problem.

Subsequently, the review delves into a literature review of generalization to different magnification levels, unseen labels, and OOD target data. The aim here is to offer a

comprehensive overview of these significant areas and highlight their relevance to this research.

Finally, the chapter concludes the literature review by assessing works that emphasize invariant features and those that leverage domain-specific features. This chapter seeks to illuminate the importance of these techniques and their potential to contribute to the development of robust, reliable, and generalizable models in digital histopathology.

Through this extensive literature review, the chapter aims to offer a comprehensive backdrop against which this research unfolds, contextualizing this thesis work within the broader scope of the field.

2.1 Digital Histopathology

The advent of digital histopathology has been marked by considerable evolution in image acquisition methods, shifting from conventional camera-based static acquisitions to the relatively novel technique of WSI [58]. Also referred to as "virtual microscopy," WSI replicates the process of light microscopy using computer technology, thus enhancing accessibility and convenience for pathologists [169]. The advent of affordable storage solutions and high-speed networks have further bolstered the efficacy of digital slide images, rendering them easily manageable.

Emerging at the intersection of technology and pathology, WSI facilitates the viewing of digitized histological slides on computer screens. It enables the seamless application of advanced image analysis algorithms, paving the way for intuitive and comprehensive slide examination. Owing to its capacity for quantifiable tissue analysis, histopathology image analysis has emerged as the gold standard for cancer recognition and diagnosis [77, 240, 101, 40]. This technology provides valuable support to pathologists, facilitating more accurate diagnosis through precise quantitative analysis.

The digital representation of a pathology slide involves a series of successive steps [1]. The process begins with fixation in formalin, which serves to prevent deterioration. Following this, the sample is embedded in paraffin, enabling the slicing of thin sections using a microtome. These wafer-thin sections are then colored using standard stains, such as H&E, transforming them into shades of pink and purple. This color transformation aids in the identification of any pathological changes under the microscope. These stained sections are subsequently scanned to digitize them at high magnifications using advanced scanners (e.g., 20x or 40x). This results in high-resolution images with dimensions varying from 10,000 to 100,000 pixels. Pathologists analyze these extraordinarily large images at multiple magnification levels to gain a comprehensive understanding of the specimen [47]. Higher magnification levels allow for the examination of smaller areas in greater detail, enabling experts to observe minute tissue characteristics that are crucial for diagnosis.

With the exponential growth in the volume of WSIs, concerted efforts have been directed towards their analysis using machine learning-based digital image analysis. This

innovative approach aids pathologists in various tasks, including diagnosis [181]. Furthermore, TCGA project has significantly enriched the availability of digital H&E WSIs, making them publicly accessible [34, 78]. However, a significant challenge lies in the fact that these images lack annotations. This underlines the need for continued research and development efforts to enhance the utility and applicability of digital histopathology in the diagnosis and treatment of diseases.

2.2 Representation Learning

Representation learning, a pivotal process in ML, involves extracting compact, informative feature vectors from raw data, thereby simplifying the task of developing classifiers or other predictive models. A high-quality representation is typically one that encapsulates the posterior distribution of the fundamental explanatory factors inherent to the observed input [19]. Such potent and expressive representations then serve as the input to supervised predictive models, contributing to their performance and accuracy.

Among the myriad methodologies for learning representations, this study is centered on deep learning approaches. Deep learning methods leverage multiple non-linear transformations, resulting in more abstract and ultimately more efficient representations. These methods underscore the role of abstraction in discerning relevant patterns and correlations, thereby enhancing the functionality and proficiency of representation learning.

The advent of CNN marked a significant stride forward in the field of representation learning. CNN brought forth the possibility of an integrated framework for simultaneous representation learning and classification, thereby streamlining the process and bolstering the model's efficiency [121]. This end-to-end approach provided by CNN revolutionizes the way we conduct representation learning, enabling more sophisticated and effective models for a wide range of tasks.

2.2.1 Convolutional Neural Networks

CNN [129] epitomize a significant stride in deep learning, owing their inception to the study of the cat's visual cortex [63]. They are a class of deep neural networks which assign varying importance, through learnable weights and biases, to different objects within an image. This assignment is achieved by learning feature maps through the utilization of multiple convolutional layers[1], as exemplified in Fig. 2.1. Convolution involves the application of a filter to the input, yielding latent space representations of images when repeated, which are then used for prediction (Fig. 2.1).

Unlike conventional hand-crafted methods, CNN provides an integrated framework for simultaneous representation learning and classification. However, due to their vast

[1]Although machine learning libraries implement cross-correlation, they refer to it as convolution [74]

quantities of adjustable parameters, CNNs are data-intensive. Several CNN architectures have been proposed, including LeNet [130], AlexNet [121], VGG [211], ResNet [84], and DenseNet [97], to name a few [222].

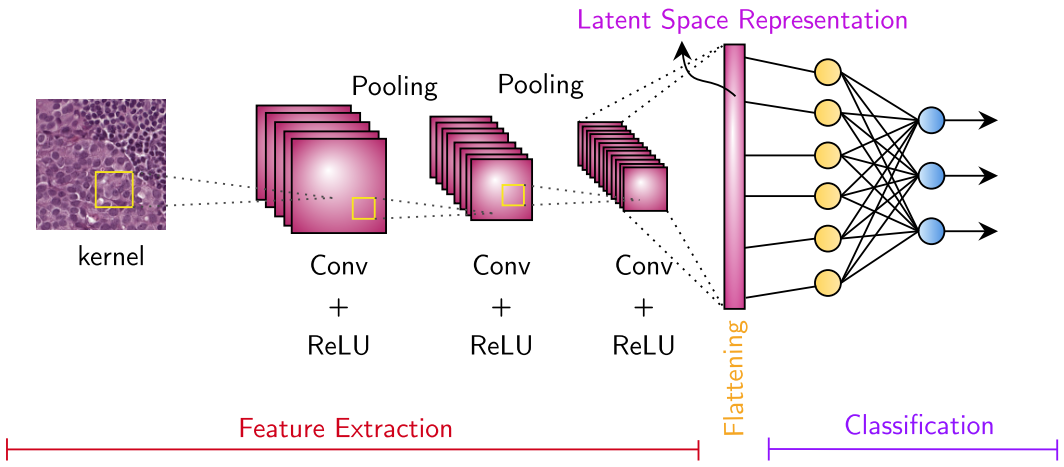


Figure 2.1: An example of a CNN architecture. Convolutional layers, pooling layers, and fully connected layers are the building components of a typical CNN architecture. Each convolution layer and pooling layer is followed by one or more fully connected layers in the typical architecture. Forward propagation is the process of transforming input data into output data.

AlexNet

Characterized by its 60 million parameters and 8-layer structure (which includes 5 convolutional layers, accompanied by 3 pooling layers, and 3 fully-connected layers), AlexNet [121] was the victorious entry in the ILSVRC 2012 competition. As depicted in Fig. 2.2, the architecture of AlexNet does not incorporate any skip connections. AlexNet extended the foundational LeNet [130] and drastically outperformed the hand-crafted approaches on the ILSVRC 2012 challenging dataset. While newer, more efficient network architectures have since been proposed, AlexNet, particularly its ImageNet pre-trained version, remains a preferred choice as a feature extractor when network architecture is not a primary concern [190, 166, 140].

ResNet

ResNet, short for Residual Network [84], marked a turning point in deep learning. As shown in Fig. 2.2, it introduced the concept of skip connections or shortcut connections,

enabling the training of exceptionally deep neural networks. ResNet reformulates layers as learning residual functions relative to the layer inputs, addressing the degradation problem that arises with increased network depth. The result is a network with over 100 layers, exhibiting no degradation in performance. Given its impressive performance and simplicity, ResNet models have found widespread use in various applications and laid the groundwork for many subsequent architectures [97, 253].

DenseNet

Depicted in Fig. 2.2, DenseNet [97] or Dense Convolutional Network, represents a different approach to layer connections. Unlike ResNet's selective layer inputs, DenseNet connects each layer to every other layer in a feed-forward fashion. By improving the information flow between layers, DenseNet reduces the number of required parameters compared to traditional convolutional networks, obviating the need to relearn redundant feature maps. Consequently, DenseNet significantly diminishes the vanishing-gradient problem, fortifies feature propagation, encourages feature reuse, and drastically reduces the number of parameters, making it an efficient and effective network for a broad spectrum of applications.

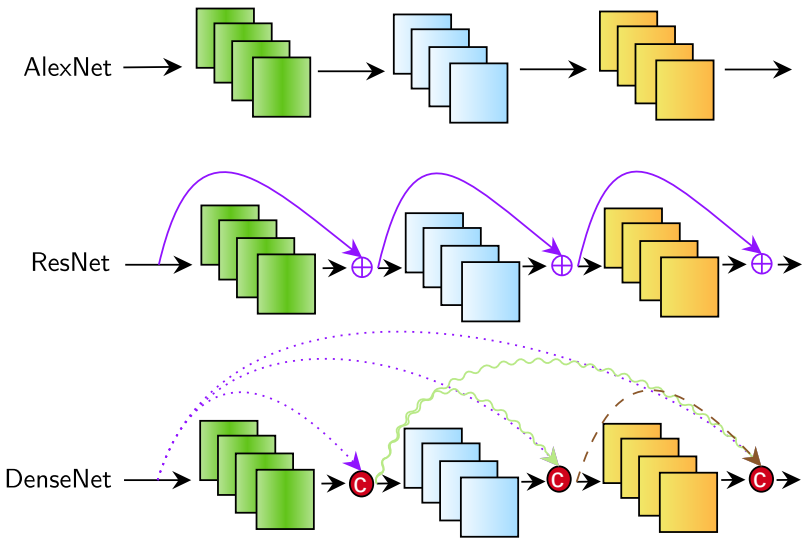


Figure 2.2: Contrasting architectures of various convolutional neural networks. The symbol $\oplus$ stands for element-wise addition and ⓒ stands for channel-wise concatenation.

Newer deep learning architectures have emerged, building on and refining the designs

of predecessors like AlexNet, ResNet, and DenseNet [253, 96]. One significant innovation is the Vision Transformer (ViT) [51], which, unlike CNN-based models, applies Transformer architecture from natural language processing. ViT treats an image as a sequence of patches, processed independently through transformer layers, allowing for a global contextual understanding. This marks a fundamental shift in approach, enhancing accuracy and efficiency in image classification.

2.2.2 Deep Metric Learning

The journey towards representation learning via a DNN sets off with a clear-cut objective. Typically, one might aim at diminishing the misclassification rate using the labels derived from a training dataset. In our pursuit of this objective, we come across a plethora of functions [103] that could be potentially leveraged. However, not all objectives are confined to reducing misclassification rates. A distinctive category of objective functions focuses on the learning of similarity or dissimilarity measures across pairs or triplets of data points, an approach commonly termed as Deep Metric Learning (DML). Research has suggested [109, 218] that utilizing these types of objective functions can culminate in a more discriminative latent space representation, thereby enhancing the performance of the learning algorithm. By honing the acuity of the discriminative capability of our model, we aim to forge a more robust and reliable representation learning method.

Siamese Networks

Siamese networks, originally introduced by Bromley et al. [26], are a specialized type of DNN architecture that comprise multiple subsidiary networks, also known as 'backbone' networks, that share weights. The prime strength of these networks lies in their capability to gauge the similarity or disparity between pairs or triplets of input data points, making them particularly effective for tasks such as signature verification or face recognition.

Training a Siamese network involves the use of specialized objective functions. Two such eminent loss functions that have gained widespread use in the realm of Siamese networks are the Contrastive loss [79] and Triplet loss [193]. The Contrastive loss is generally employed when the Siamese network comprises two sub-networks, while the Triplet loss function is adopted when the Siamese network features three sub-networks.

For understanding these loss functions, let's denote the anchor, positive, and negative samples in a triplet as x_a, x_p, and x_n, respectively. The Contrastive loss and the Triplet loss functions can be formally defined as:

$$\ell_{\mathrm{c}} = \sum_{i=1}^{b} \left[(1-y)||\mathbf{f}(x_1^i) - \mathbf{f}(x_2^i)||_2^2 + y \left[-||\mathbf{f}(x_1^i) - \mathbf{f}(x_2^i)||_2^2 + \zeta \right]_+ \right], \tag{2.1}$$

$$\ell_{\mathrm{t}} = \sum_{i=1}^{b} \left[\|\mathbf{f}(x_a^i) - \mathbf{f}(x_p^i)\|_2^2 - \|\mathbf{f}(\mathbf{x}_a^i) - \mathbf{f}(x_n^i)\|_2^2 + \zeta \right]_+, \tag{2.2}$$

In these equations, $\mathbf{f}(.)$ symbolizes the output of the network, $\zeta > 0$ is a margin that helps to separate the positive and negative pairs, y is a binary variable which equals zero when the samples originate from the same class and one when they hail from different classes, b denotes the mini-batch size, $[\cdot]_+ := \max(\cdot, 0)$ symbolizes the standard Hinge loss, and $||\cdot||_2$ denotes the ℓ_2 norm. These loss functions essentially strive to shape the network in such a way that it learns to embed samples from the same class close together and samples from different classes far apart in the feature space, leading to a more discriminative latent space representation.

2.2.3 Transfer Learning: Overcoming the Bottleneck

The primary challenge or bottleneck in leveraging deep neural networks, particularly CNNs, is their voracious appetite for large volumes of labeled data. The larger the amount of data fed into these models, the better they learn to discern patterns and make accurate predictions. This relationship has been explored in depth by Sun et al. [220], who discovered that the performance of CNNs on visual tasks follows a logarithmic growth pattern in relation to the size of the training dataset.

However, in certain fields such as the medical domain, obtaining large, well-annotated datasets for training purposes can be a formidable task. This is primarily due to the sensitive nature of medical data, stringent privacy regulations, and the necessity for expert annotation, which is both time-consuming and expensive.

To circumvent this limitation, the strategy of transfer learning has been employed [234]. Transfer learning involves pre-training a model on a large, readily available dataset and then fine-tuning it on the task-specific, often smaller, dataset. The underlying premise is that the features learned by the model in the pre-training phase, such as basic shape and pattern recognition in the case of image data, can be repurposed and refined to perform well on the target task. This approach alleviates the necessity for colossal amounts of labeled data, making it a valuable technique in domains where such data is scarce or difficult to procure.

2.3 Transfer Learning

Transfer learning is a ML approach in which a model that has been trained on one objective is re-purposed for a second related task (see Fig. 2.3). In transfer learning terminology,

the target-domain task is the ultimate task and the source-domain task is a related task or some related tasks which will facilitate solving the target-domain task.

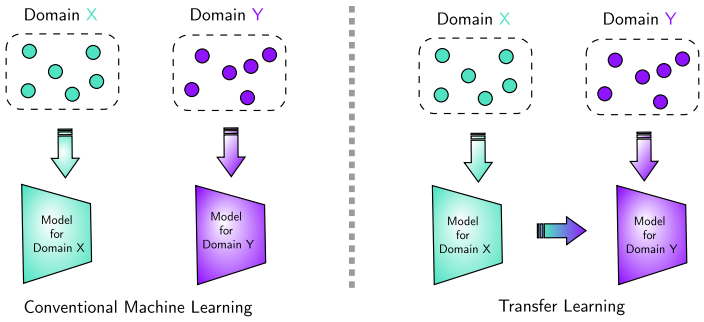


Figure 2.3: As illustrated on the left, traditional machine learning constructs each model on a specific domain in isolation; however, with transfer learning (right), the target model is created using the learned knowledge from the source-domain.

Transfer learning has been based upon an important assumption as follows. Although multiple different tasks (source-domain and target-domain tasks) may be i.i.d. but, deep inside, they share a *common low-dimensional representation* [56]. One way to perform transfer learning is to learn a shared latent space representation across related tasks. The neural codes, commonly known as DeCAF features [50], consist of reusing existing ImageNet pre-trained [235] CNNs (architecture and parameters) to perform target-domain task, using retraining only the fully-connected layers [50, 6, 201]. Bar et al. [10] showed that DeCAF features might be a suitable substitute for domain-specific representations for general medical domain image recognition tasks. Lu et al. [147] fine-tuned a pre-trained AlexNet [121] to classify a short-size Magnetic Resonance Imaging (MRI) dataset.

Transfer learning has also been found to be promising for histopathology images [113]. Xu et al. [254] showed that DeCAF features can be representative enough for MICCAI 2014 Brain Tumor Digital Pathology challenge achieving 97.5% classification accuracy. Spanhol et al. [213] used DeCAF features for breast cancer histopathological image recognition task and demonstrated that DeCAF features are superior to hand-crafted features.

Some works have used pre-trained CNNs but through domain-related tasks. For instance, Faust et al. [60] used a brain-tumor-educated CNN to perform a classification on RCC. Bayramoglu et al. [11] analyzed and showed the effectiveness of transfer learning compared to learning from scratch for cell nuclei classification. Riasatian et al. [180] utilized a TCGA fine-tuned densely-packed CNN and demonstrated that their fine-tuned network using patches extracted from TCGA can outperform DeCAF features in histopathological image search and classification tasks.

It may be prohibitively expensive and time-consuming to train a new DNN for each new application without the ability to transfer relevant knowledge from prior datasets. The absence of annotated labels with varied and generalizable tissue types from different

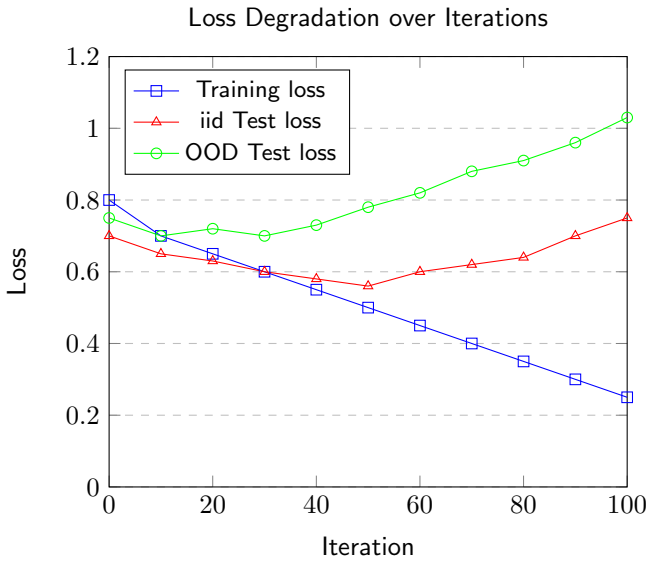


Figure 2.4: The (blue) training, (red) iid validation, (green) OOD test losses vs. iterations.

organs, as well as the diversity in WSI scanners and staining methods, are two major barriers to transfer learning. On top of that, sometimes a DNN model trained on the source-domain task(s)[2] cannot generalize to perform well on the target-domain task[3]. The principal reason is *domain shift* which will be discussed in Section 2.3.2. In the subsequent subsection, the generalization problem will be discussed in more detail.

2.3.1 The Problem of Generalization

At its core, ML is fundamentally concerned with the problem of generalization. One of the central questions that arise in this field is how we can ensure that the models we train will perform adequately when faced with new, unseen data. This issue is especially pertinent in the realm of deep learning, where the vast weight space of DNN models can easily lead to overfitting the training data, consequently compromising generalization.

The challenge intensifies significantly when the unseen data is characterized by a distinct distribution, a scenario known as domain shift. In this situation, even the most minor changes in the statistical properties of the data compared to the training data can lead a DNN model to fail. Some essential considerations for training a generalized DNN include:

- **Dataset Diversity**: To train a generalized DNN, a diverse dataset is imperative. By diversity, we are not alluding to sheer volume; instead, we refer to a dataset that embodies a wide variety of samples. This breadth of data ensures the DNN model

[2]In medical literature It is also called the *internal dataset(s)*.
[3]In medical literature It is also called *external validation dataset*.

is exposed to more than just a specific subset of data, thus paving the way for more effective generalization.

- **Model Complexity**: ML models that become excessively complex are prone to overfitting, while simpler, more compact models tend to mitigate this risk. Numerous methods in machine learning, known as regularization methods, discourage the learning of more complex models to avert overfitting.

- **Regularization**: Regularization comprises a suite of techniques that reduce the complexity of a DNN model during training, thereby curtailing overfitting [183]. This process can involve various strategies, including the use of dropout and L_1-norm or L_2-norm to regularize the cost function.

While adhering to these considerations can foster a more generalized DNN, a DNN model may still struggle with distribution shifts, impeding effective generalization to unseen, target-domain data. As illustrated in Fig. 2.4, although using a hold-out validation set can decrease the loss for the target-domain task, the loss can increase at some point before reaching optimal iterations. This situation demonstrates the persistent challenge of achieving true generalization in the context of distribution shifts.

2.3.2 Understanding Distribution Shift

The concept of distribution shift refers to the divergence between source-domain and target-domain data distributions [175]. In an ideal scenario, where distribution shift is absent, the knowledge obtained from the source domain can be directly transposed to the target domain [61]. However, real-world scenarios often differ markedly from this idealized situation.

A prevalent assumption in most statistical learning methods is that the source and target domain data sets are i.i.d., while the frequently occurring OOD scenarios are often overlooked. Under this assumption, the focus is primarily on minimizing errors within the source domain, anticipating that this will yield similar success in the target domain. However, this assumption can lead to significant misestimations, as even the slightest divergence in the statistical properties of the target domain data set compared to the source domain can precipitate a catastrophic failure of the method [233].

In response to these challenges, MDL approaches have gained prominence. These approaches come into play when the presumption that the "source-domain and target-domain derive from a nearly identical distribution" does not hold *true*. MDL thus strives to address this discrepancy, working to ensure a model's resilience and effectiveness in the face of distribution shifts, thereby enhancing its capacity to generalize across diverse data sets.

2.4 Multi-Domain Learning

In MDL, the aim is to train a single model which is effective for multiple known domains [43, 23]. There are two variants of MDL in the literature that can be confused, i.e., DA and DG; there are some nuances that make them different from each other. In the next section, these nuances are discussed in detail.

2.4.1 Domain Adaptation

DA techniques have been developed as solutions to mitigate the problem of domain shift, typically achieved by aligning the target and source-domain distributions within a domain-invariant feature space [16]. These techniques usually presuppose access to a few [124] or unlabelled [66, 89, 236, 188] samples from the target domain.

According to Farahani et al., [57], DA methods can be categorized into four types based on the category gap: closed-set, open-set, partial, and universal DA.

- **Closed-set DA**: In this scenario, both the target and source-domain datasets share the same classes, but there is a domain gap between the domains. Despite sharing the same categories, the distribution of features within these categories may differ significantly, presenting a challenge for effective adaptation.
- **Open-set DA**: Here, both source and target-domain datasets, in addition to common labels, might have unique class labels [167]. This adds a layer of complexity as the model must not only adapt to new feature distributions but also potentially unseen categories.
- **Partial DA**: This refers to the situation where the target label set is a subset of the source label set [30, 263]. In this case, the model must learn to identify and focus on relevant categories while ignoring extraneous ones.
- **Universal DA**: This most challenging form generalizes all the above scenarios and is not restricted to any prior knowledge regarding the source and target-domain dataset labels [260]. It requires models to be highly flexible and capable of adjusting to multiple forms of domain shifts.

Within the realm of histopathology, domain-invariant techniques are commonly developed by incorporating a domain adversarial module into the model. For instance, Aubreville et al. [21] trained and evaluated mitosis detection models that included a domain adversarial module to assist with generalization on canine cutaneous mast cell cancer and canine mammary carcinoma. Another study by Lafarge et al., [126] compared traditional augmentation and normalization methods with domain-adversarial neural networks (DANN) as an alternative approach. Chang et al., [33] proposed the stain mix-up method,

an effective data augmentation strategy for differentially stained histopathology images. This approach encourages the model to adapt to varied stain colors in an unsupervised manner. In all these techniques, the data from the target domain are accessible, though the labels or annotations may not be available.

2.4.2 Domain Generalization

The concept of DG represents a more formidable challenge within the spectrum of DA problems. The key objective under the purview of DG is to attain a level of generalization that allows for accurate predictions on a completely unseen or novel domain. This is achieved by training on a multi-domain source-domain dataset, without the need to retrain the network with data from a held-out domain. This reference to the 'held-out domain' implies a portion of the data that has been purposely set aside and not included in the training process.

The premise of DG is highly relevant and practical, often regarded as more realistic when compared with DA. The basis of this comparison stems from the fact that, in many real-world scenarios, we are often faced with situations where we have to make decisions or predictions based on new, unseen data. More importantly, we usually do not have any prior knowledge or information about the distribution characteristics of this new target domain data set.

Considering that there is no prior knowledge about the target-domain distribution, a key question that emerges within the context of DG pertains to the model's guidance strategy. More specifically, how can we instruct or navigate the model to extract information from the source data in such a way that it is not just discriminative, but also resilient to changes in domain distribution? This robustness to domain variations is a fundamental requirement for the model to ensure its performance remains consistent even when confronted with new data that exhibit different characteristics from the training data.

The field of DG is generally divided into three primary categories. These categories, namely (1) Selecting, (2) Domain-Invariant-Based, and (3) Hybrid Domain-Invariant and Domain-Specific-Based, each adopts a unique approach to address the challenge posed by domain shifts.

Selecting methods

The first category, termed "Selecting", operates on a model-specific basis for each domain within the source-domain dataset. This strategy essentially involves training individual models for every available domain and then choosing the most relevant model when presented with a new target domain [256]. This relevance is typically established by comparing the inherent characteristics of the target domain with the source domains, and the model associated with the most similar source domain is selected for application on the target domain.

Domain-Invariant-Based methods

These methods [112, 161, 69, 161, 65, 25] embrace a more holistic approach, predicated on the belief that every domain consists of a universal, or common component, and a unique, or domain-specific component. The strategy here is to isolate the common elements that pervade numerous domains and utilize this to cultivate a durable, domain-invariant feature representation. This commonality essentially represents shared or overlapping information across disparate domains. Extracting these common elements facilitates the creation of a representation that remains primarily unaffected by domain-specific changes, thereby enhancing the model's capacity to generalize effectively to unseen domains.

One promising direction to augment the generalization capabilities of machine learning models is through MDL, with a specific emphasis on learning domain-invariant features. This method involves simultaneous processing and learning from multiple source domains. As a result, a model can obtain a more comprehensive and diverse representation of data, which aids in more effectively capturing the inherent structure of the data manifold.

The focal point here is the learning of domain-invariant features - those aspects consistent across all domains, unaffected by specific characteristics of any single domain. These invariant features can be perceived as common threads across domains, encapsulating shared information, and thereby effectively decoupling task-related knowledge from the specificities of any single domain.

By concentrating on these core, task-specific features rather than the nuances of individual domains, the model becomes more resistant to domain shift. This approach encourages robust generalization and enhances the model's performance on unseen domains. It diminishes the risk of overfitting to any particular source domain and increases the model's overall flexibility and adaptability. This strategy offers a potent tool for effectively handling the inherent diversity and variability in real-world data, enabling the model to better handle new, unseen domains.

A promising research direction in DG is to utilize the learning regime of Model-Agnostic Meta-Learning (MAML) [62]. MAML-based DG frameworks divide the source-domain dataset into meta-train and meta-test domains to simulate the domain shift [134]. This approach has shown promise in terms of generalization and fast adaptation [243, 172, 7]. The learning goal of MAML-based DG methods is to update a model using the meta-source-domain(s) in a way that minimizes the test error on the meta-target-domain, which is typically achieved via bi-level optimization.

In the field of histopathology, MAML-based DG has found extensive use. For example, Cai et al. [28] introduced a DG technique for multiclass recognition of nucleus in Ki67 IHC images, aimed at addressing practical limitations of DG, including the insufficient number of domains and the issue of class mismatching across domains.

While there are many DG approaches available, Dou et al. [52] proposed an MAML-based DG technique, called MASF, that aims to learn the semantic structure of the source-domain dataset(s) while also promoting class-specific cohesiveness and separation across

domains using DML approaches. MASF represents an important step in DG research, as it emphasizes not only domain-invariance but also class-specific discriminability, making it a valuable tool for the practical application of DG. MASF follows MAML learning regime; thus, in what follows, MAML will be reviewed first.

Model-Agnostic Meta Learning– MAML (a.k.a. *learning to learn*) [62] is a heated topic in ML with the main idea of "the small number of gradient steps along with a small amount of training data from a new task are sufficient to achieve strong generalization on that specific task".

Algorithm 1: MAML Algorithm

$p(\mathcal{T})$: distribution over tasks
α and β: step size hyper-parameters
$\theta \leftarrow \theta_0$;
while *not done* **do**
  sample a batch of tasks: $\mathcal{T}_i$
  for *all* $\mathcal{T}_i$ **do**
    Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}$
    Compute adapted parameters with gradient descent:
    $\theta_i' \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$,
  end
  Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}^{(1)}(f_{\theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}^{(0)}(f_\theta)})$,
end

Consider a model that is represented by a parametric function f_θ. When the model is adapted to a new task $\mathcal{T}_i$, the parameters of the model become θ_i'. The updated parameter is produced by performing one or more gradient descent updates on the task. For instance, when merely one gradient update is used,

$$\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}\left(f_\theta\right). \tag{2.3}$$

The model parameters are trained by optimizing for the performance of $f_{\theta'}$ across tasks sampled from $p(\mathcal{T})$. To put it another way, the meta-objective is as follows:

$$\min_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}\left(f_{\theta_i'}\right) = \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}\left(f_{\theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)}\right). \tag{2.4}$$

Where: - θ represents the model parameters to be optimized. - $\mathcal{T}_i$ is a task sampled from the distribution $p(\mathcal{T})$. - θ_i' denotes the modified model parameters for task $\mathcal{T}_i$. - $f_{\theta'}$ represents the model with parameters θ'. - $\mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'})$ is the loss on task $\mathcal{T}_i$ using model $f_{\theta_i'}$. - α is the step size for the inner update. - $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$ is the gradient of the loss on task $\mathcal{T}_i$ with respect to θ.

One has to note that the meta-optimization is conducted on the model parameters, whilst the objective function is derived using the modified model parameters. In fact, MAML aims to optimize model parameters in such a manner that a single or a small number of gradient steps on a new task produces the most possible appropriate behavior. The meta-optimization across tasks is carried out using Stochastic Gradient Descent (SGD), with the following model parameters updated:

$$\theta \leftarrow \theta - \beta \nabla_{\theta} \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i} \left(f_{\theta_i'} \right). \tag{2.5}$$

Where: - β is the learning rate for the outer update. - $\nabla_{\theta} \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i} \left(f_{\theta_i'} \right)$ is the gradient of the sum of losses across tasks with respect to θ.

According to Algo. 1, The Eqs. (2.3, 2.5) are repeated iteratively until convergence.

Model-Agnostic learning of Semantic Features– Inspired by the MAML approach, MASF learns a latent space representation suitable for generalization to an unseen target-domain. MASF is a model agnostic learning scheme that can be used for performing different tasks. If the classification is the task at hand, MASF consists of some layers of feature extraction, classification, and metric embedding.

Consider that a CNN, consisting of some feature extraction and classification layers, is being used as the model in a MASF framework. Then, G_{ψ} sub-network is the feature extractor which ends up in latent space representation, and S_{θ} sub-network is the classification layers of the model of this CNN. A metric learning module, M_{ϕ} also comes after G_{ψ} which can be a triplet loss using Eq. 2.2, or contrastive loss functions using Eq. 2.1.

MASF combines three different loss functions for learning a discriminative embedding space. These loss functions are a cross-entropy loss for supervised learning, a Kullback–Leibler (KL) divergence loss between pairs of meta-train and meta-test sets for domain alignment purposes, and a metric loss to promote domain-independent class-specific cohesion and separation of instances.

At each iteration of the algorithm, a batch of the source-domain dataset ($\mathcal{D}$) is split into meta-train and meta-test batches, indicating by $\mathcal{D}_{\text{tr}}$, and $\mathcal{D}_{\text{te}}$, respectively. Then, G_{ψ}, S_{θ}, and M_{ϕ} sub-networks are updated using three aforementioned loss functions as below: The first loss which is used is the cross-entropy loss for supervised learning:

$$\mathcal{L}_{\text{ce}}(\mathcal{D}_{\text{tr}}; \psi, \theta) := \frac{-1}{|\mathcal{D}_{\text{tr}}|} \sum_{\mathcal{D} \in \mathcal{D}_{\text{tr}}} \frac{1}{|\mathcal{D}|} \sum_{(x} , y) \in \mathcal{D} \sum_{c=1}^{C} \mathbb{I}(y = c) \log \mathbb{I}(\widehat{y} = c), \tag{2.6}$$

where $|\cdot|$ denotes the cardinality of a set, C is the number of classes, y is the true label for x, $\widehat{y}$ is the predicted label for x, and $\mathbb{I}(\cdot)$ is the indicator function which is one when its condition is satisfied and is zero otherwise. In practice Eq. 5.2 is used in batches. The updated ψ and ϕ are denoted by the cross-entropy loss by ψ' and ϕ', respectively. The cross-entropy loss function is used for hard class separation.

The next loss used is a KL-loss which takes care of between-domain generalization. Consider the feature extraction embedding of samples of domain k. If the samples of this domain in the label c are denoted by $\{x_{c,i}^{(k)}\}_{i=1}^{n_c}$ and their feature extraction embeddings are $\{G_{\psi'}(x_{c,i}^{(k)})\}_{i=1}^{n_c}$, the Monte-Carlo approximation for the mean embedding of label c in domain k can be given as

$$z_c^{(k)} := \frac{1}{n_c} \sum_{i=1}^{n_c} G_{\psi'}(x_{c,i}^{(k)}). \tag{2.7}$$

This mean is passed through the layers for supervised learning and its softmax with temperature $\tau > 1$ is

$$s_c^{(k)} := \text{softmax}(S_{\theta'}(z_c^{(k)})/\tau), \tag{2.8}$$

which provides a soft confusion matrix. For two domains $\mathcal{D}_i$ and $\mathcal{D}_j$, the symmetrized KL divergence, averaged over all the C classes, is calculated as

$$\ell_{\text{domain alignment}}(\mathcal{D}_i, \mathcal{D}_j; \psi', \theta') := \frac{1}{C} \sum_{c=1}^{C} \frac{1}{2} \Big[D_{\text{KL}}(s_c^{(i)} \| s_c^{(j)}) + D_{\text{KL}}(s_c^{(j)} \| s_c^{(i)}) \Big], \tag{2.9}$$

where D_{KL} denotes the KL divergence. This loss is computed over all the meta-train and meta-test trial sites through

$$\mathcal{L}_{\text{domain alignment}}(\mathcal{D}_{\text{tr}}, \mathcal{D}_{\text{te}}; \psi', \theta') := \frac{1}{|\mathcal{D}_{\text{tr}}||\mathcal{D}_{\text{te}}|} \sum_{\mathcal{D}_i \in \mathcal{D}_{\text{tr}}} \sum_{\mathcal{D}_j \in \mathcal{D}_{\text{te}}} \ell_{\text{gen}}(\mathcal{D}_i, \mathcal{D}_j; \psi', \theta'). \tag{2.10}$$

In practice, we use Eqs. (2.7, 2.9, and 2.10) on batches. This KL-divergence loss function is used for domain generalization between each pair of domains.

Finally, a metric loss [193] is used for the sake of promoting domain-independent class-specific cohesion and separation of instances. Consider a triplet of anchor, positive, and negative instances denoted by x_a, x_p, and x_n, respectively. Triplet loss attempts to increase the inter-class variances of data and decrease the intra-class variances. If we randomly sample R triplets, $\mathcal{T} := \{x_a^r, x_p^r, x_n^r\}_{r=1}^{R}$, from all the source-domain datasets $\{\mathcal{D}_k\}_{k=1}^{K}$, the average triplet loss is

$$\begin{aligned} \mathcal{L}_{\text{triplet}}(\mathcal{T}; \psi', \phi) := \frac{1}{R} \sum_{r=1}^{R} \Big[& \|M_\phi(G_{\psi'}(x_a^r)) - M_\phi(G_{\psi'}(x_p^r))\|_2^2 \\ & - \|M_\phi(G_{\psi'}(x_a^r)) - M_\phi(G_{\psi'}(x_n^r))\|_2^2 + \zeta \Big]_+, \end{aligned} \tag{2.11}$$

where $\|\cdot\|_2$ is the ℓ_2 norm, ζ is a margin and $[\cdot]_+ := \max(\cdot, 0)$ is the standard Hinge loss. The triplet loss function is used for metric learning and soft class separation where the intra-class and inter-class variances are decreased and increased, respectively.

The MASF learning scheme is characterized by an episodic training scheme derived from MAML [62], which exposes the model optimization to distribution mismatch. To achieve domain generalization, the model is trained on a succession of simulated domain shifts in each episode. To be more precise, each iteration randomly divides the available source-domain datasets, i.e. $\mathcal{D}$, into sets of meta-train $\mathcal{D}_{\text{tr}}$ and meta-test $\mathcal{D}_{\text{te}}$ domains. After optimization with one or more steps of gradient descent using $\mathcal{D}_{\text{tr}}$, the model learns to perform well semantically on hold-out $\mathcal{D}_{\text{te}}$.

MASF starts with updating the parameters of the G_ψ and S_θ, in which ψ and θ are first adjusted using a cross-entropy loss function $\mathcal{L}_{\text{ce}}$ on the classification task, which is calculated on $\mathcal{D}_{\text{tr}}$ using Eq. 5.2. After obtaining this updated G_ψ and S_θ subnetworks, a meta-learning phase is performed to enforce particular model attributes on the hold-out domain, i.e. $\mathcal{D}_{\text{te}}$, using the domain alignment loss function, using Eq. 2.10. Using the domain alignment loss function, parameters are changed in such a way that future updates with provided $\mathcal{D}$ increase the model's generalizability for unseen target-domain(s). On top of that, MASF tries to cluster the datasets compactly according to class labels regardless of the domain using a metric loss function, as described in Eq. 5.3, and through this loss function M_ϕ subnetwork is updated. In this work, the MASF method will be used for developing a hospital-agnostic model.

Hybrid Domain-Invariant and Domain-Specific-Based methods

The final category of domain generalization methods incorporates a hybrid approach that appreciates the unique characteristics present in each domain while still emphasizing the extraction of domain-invariant features. This approach deviates from solely focusing on shared commonalities across domains, emphasizing instead the value of domain-specific attributes. Such unique features across various domains can provide critical insights into the diversity of data, thus enabling models to yield a more customized performance when encountering a novel domain.

This hybrid strategy operates on the principle of balance. It aims to intertwine domain-invariance with domain-specificity in such a way that maximizes the adaptability and resilience of the model to different domains. The value of this equilibrium becomes evident when considering the pitfalls of completely disregarding domain-specific information. Despite the appealing qualities of domain-invariant features, especially in their role in building models resistant to domain shift, over-reliance on them may not ensure the best generalization performance. This viewpoint has been emphasized in the research conducted by Mancini et al. [152] and Shankar et al. [200], where they highlight that essential, discriminative information could be missed if domain-specific characteristics are overlooked.

Bui et al. [27] took note of this and proposed the mDSDI method, aiming to harmonize the use of domain-specific information and domain-invariant features. The mDSDI method extracts and utilizes both invariant and unique features across various domains, fostering a more comprehensive and potentially more effective representation of data. Their mathematical proof lends robust support to the efficacy of this approach, further underscoring

its potential as a refined solution to domain generalization challenges. By embracing the strengths of both domain-invariant and domain-specific features, this hybrid method represents a promising strategy in the domain generalization landscape.

Meta-Domain Specific-Domain Invariant (mDSDI) Algorithm– The mDSDI algorithm is an intricate and ingenious solution aimed at ameliorating generalization performance in machine learning tasks. Fundamentally, the framework encompasses several components, each undertaking a pivotal role in the execution of the algorithm.

Primarily, there is a domain-invariant representation, denoted by Z_I, which is generated by an encoder. Additionally, there's a domain-specific representation, referred to as Z_S, created by another encoder. These representations bear great significance as they carry the domain-invariant and domain-specific information of the input data respectively. The extraction of Z_I is accomplished using an adversarial training framework. In this study, a domain discriminator attempts to classify the domain based on Z_I. Concurrently, the encoder for Z_I aims to make this task as challenging as possible. This culminates in the production of a robust domain-invariant representation. Analogously, for capturing domain-specific nuances, a domain classifier is trained to predict the domain label from Z_S.

A paramount aspect of the mDSDI methodology is the disentanglement of Z_I and Z_S, thereby safeguarding the uniqueness of the information each of them carries. This disentanglement is actualized by minimizing the covariance between Z_I and Z_S, effectively ensuring they contain non-overlapping information. The adequacy of these two representations in the context of the classification task is assured by another classifier that utilizes the combination of Z_I and Z_S to predict the original sample label. The aim is to ensure that the combined representations are sufficient to recreate the original data label. The novelty in the mDSDI methodology stems from the inclusion of a meta-learning framework for domain-specific information. This allows Z_S to adapt the information it learns from the source domains to unseen target domains, thereby aiming for enhanced generalization. This is achieved by dividing each source domain into meta-train and meta-test sets and optimizing the parameters accordingly.

Finally, the complete training procedure incorporates several objective functions that are associated with both the domain-invariant and domain-specific representations, their disentanglement, and the classification task itself. This combined objective function, which we denote as LA, forms the consolidated optimization target for the system. The purpose of this integrative approach is to attain a comprehensive optimization across all parameters involved in the model, resulting in superior generalization performance. Specifically, the function LA is minimized over the parameters $\theta_Q, \theta_{DS}, \theta_R, \theta_F$ and maximized over θ_{DI}. This is illustrated in the following equation:

$$\underset{\theta_Q,\theta_{DS},\theta_R,\theta_F}{\text{minimize}} \quad \underset{\theta_{DI}}{\text{maximize}} \quad LA := \lambda_{ZI} L_{ZI} + \lambda_{ZS} L_{ZS} + \lambda_D L_D + L_T \tag{2.12}$$

In the above equation, λ_{ZI}, λ_{ZS}, and λ_D are tuning parameters used to balance the

various components of the loss function. L_D is responsible for ensuring the disentanglement of domain-invariant and domain-specific representations, which effectively encapsulates the diverse information from the different domains. L_T is directly tied to the classification task. It evaluates the performance of the classifier, which operates on a concatenation of the domain-invariant and domain-specific features.

2.5 Impact of Pre-Trained Models on Generalization

DG, while extensively studied [266], remains a complex field with varying strategies yielding different results [76, 192]. A deeper dive into the intricacies of this field was conducted by Wiles et al. [250], who investigated distribution shifts in three key areas: (1) spurious correlations, (2) low-data environments, and (3) unfamiliar scenarios. Their exploration highlighted the effectiveness of seemingly simple strategies, such as data augmentation and pre-training. Nevertheless, their work also revealed that the performance of DG algorithms is dependent on the specific dataset and the nature of the distribution shift.

This context-specific effectiveness emphasizes the need for a greater understanding and improvement of domain generalization strategies, particularly to ensure their robustness in real-world scenarios. A question that naturally arises from this is whether the field has significantly advanced beyond standard ERM algorithm [250] capabilities.

While some might view these findings as discouraging, it's essential to note that there is research demonstrating the potential of generalization across varying dataset distributions [250, 176]. In particular, certain studies suggest that pre-training models using large datasets can be quite effective for OOD generalization [226, 250]. In the ensuing sections, we delve into a few specific pre-trained models and their impact on generalization.

2.5.1 Vanilla Pre-Trained Models using ImageNet

Pre-training is a dominant paradigm in computer vision. It is based on the concept that if models are trained on one dataset, they would be able to provide insights into other related tasks. This has led to the prevalence of Vanilla ImageNet pre-trained models that are frequently used for various computer vision tasks [73, 50, 141, 35]. However, the effectiveness of these models has been called into question in certain contexts. For example, it was found that these models did not perform well in specific tasks such as Microsoft COCO object detection [202, 137]. In fact, a model trained with random initialization was found to outperform these models in certain scenarios [68]. This casts some doubt on the utility of relying solely on vanilla pre-trained models.

2.5.2 SSL and SWSL Pre-Trained Models

A common sentiment in the AI community is that pre-training models on more diverse datasets would lead to better OOD generalization. This is further supported by studies showing that pre-trained models using more varied datasets result in improved OOD generalization during real-life distribution shifts [226, 87]. Some pre-trained models [258] have outperformed the vanilla ImageNet pre-trained models in both OOD and in-distribution top-1 accuracy levels. Among the various methods, two particular approaches are noteworthy, both introduced by the Facebook team: the Semi-SL [258] and the Semi-WSL [258] pre-trained models.

2.5.3 KimiaNet: A Pre-Trained Model for Histopathology

For histopathology, using a model specifically pre-trained for this field could potentially yield better results than models trained on natural images. One such model is KimiaNet [180], a pre-trained model developed using the DenseNet topology [97] and trained on TCGA dataset, a diverse, multi-organ public image repository.

2.6 Summary

In this chapter, I provided an extensive review of the scholarly literature pertaining to the topics and concepts that are central to the content of this thesis. The discussion commenced with an examination of CNN and their well-known architectures, underscoring their significance in the landscape of Deep Learning (DL).

Further, the review delved into the concept of transfer learning, a method that allows us to leverage pre-existing models for new tasks, potentially saving computational resources and time. A particular emphasis was placed on MDL, a technique that has been instrumental in addressing the challenge of generalization that DNNs commonly encounter.

The discussion culminated with an exploration of the role and utility of pre-trained models. The advent of numerous such models trained on extensive and diverse datasets has provided an opportunity to study their efficacy in OOD generalization. This is particularly relevant given the ongoing debate in the AI community regarding the ability of DG to effectively tackle the generalization problem despite the complexities it introduces during the training process.

In subsequent chapters, I will explore how these methods and concepts have been applied within the field of digital pathology, including potential enhancements and a thorough analysis of the results. These sections largely constitute the product of my Ph.D. research and highlight my contributions to the field. Through these discussions, I hope to shed light on the continued progress in these areas and provide a meaningful contribution to the ongoing discourse on these topics.

Chapter 3

Generalization to Different Magnification Levels

Prologue

The content of this chapter is based on an article published during my Ph.D. research: Magnification generalization for histopathology image embedding- M Sikaroudi, B Ghojogh, F Karray, H. Tizhoosh- International Symposium on Medical Imaging (ISBI) 2020 [205]

3.1 Motivation

In recent years, the application of ML and specifically deep learning for cancer diagnosis in the histopathology field has grown exponentially [106]. However, challenges remain due to the variance in magnification levels in WSI images [195].

In the field of histopathology, a central challenge has been the discrimination of image patches, which originates from different magnification levels, to derive a more condensed representation of WSI. This task is made intricate by the prominent disparities in the visual characteristics that manifest across varying levels of magnification. Despite these patches being i.i.d.—since they all emerge from the same WSI—the differences in visual manifestations can be perceived as shifts in their distributions.

If consider the variable X to represent a patch from a WSI and M to represent a specific magnification level, one can represent the distribution of patches at a particular magnification level as $P(X|M)$.

Now, given that all patches emerge from the same WSI, one can argue that for any two magnification levels M_i and M_j, the following holds:

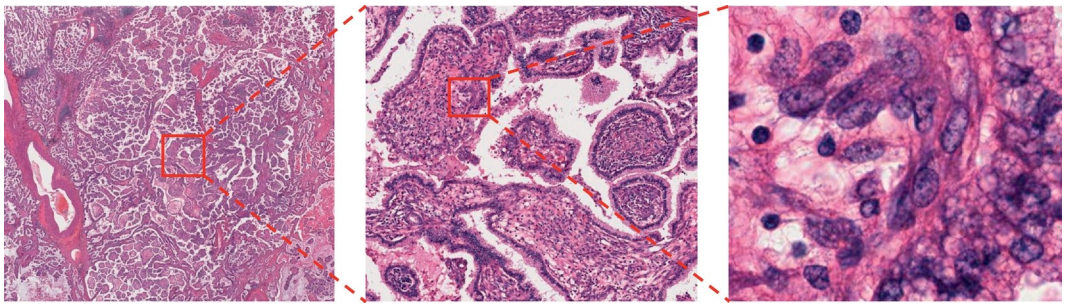


Figure 3.1: The provided WSI, sourced from TCGA, depict varying magnification levels of the same specimen. The images on the right display an enlarged view of the area highlighted by the red box in the corresponding left images. The leftmost image distinctly portrays a papillary structure, while the rightmost image provides a clear view of each cell's nucleus. Taken from [118].

$$P(X|M_i) = P(X|M_j) \tag{3.1}$$

Eq.3.1 would suggest that the patches are identically distributed across different magnification levels.However, due to differences in the visual manifestations of patches at different magnification levels, in reality, it is more accurate to say that the distributions across magnification levels are different. Therefore, it can be written as

$$P(X|M_i) \neq P(X|M_j) \tag{3.2}$$

Eq.3.2 shows that different magnification levels are indeed like parallel worlds with different "visual atmospheres," each contributing to a shift in the overall distribution of the WSI. This can be interpreted as a OOD shift between different magnification levels within the same WSI.

It's important to remember, however, that this is a simplification of a complex reality. The actual relationship between patches at different magnification levels is likely much more intricate, being influenced not only by the magnification level but also by a multitude of other factors such as the specific properties of the tissue and the particular characteristics of the imaging process. As a general guideline, according to Fig. 3.1, high-power field microscopic images proficiently capture fine details related to cell morphology, while lower-power field images provide a more comprehensive representation of structural aspects, such as the arrangement of glandular structures composed of numerous cells.

To further elaborate, imagine each magnification level as a parallel world of its own. While these worlds all belong to the same universe—similarly to how all patches come from the same WSI—they each offer unique living spaces and atmospheres. These distinct environments within each "world" or "magnification level" can be compared to the varied

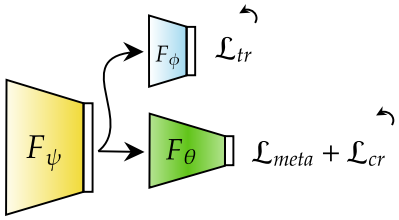


Figure 3.2: The subnetworks and loss functions used in MASF and the proposed magnification generalization for histopathology image embedding.

visual characteristics that one observes in histopathology images at different magnifications. The task, then, is to reconcile these disparate "worlds, bringing together the unique information each one provides to form a more comprehensive and compact representation of the WSI from which they originate. This challenge is the crux of this study as these parallel worlds are navigated and their unique visual atmospheres address the generalization problem in the histopathology field.

Although previous research has delved into the realm of magnification adaptation for histopathological embedding [12, 42, 215], my study embarks on a novel journey, focusing on magnification generalization, where the model in training is kept oblivious to the target magnification. This innovative line of investigation, to the best of my knowledge, has not been previously pursued.

Our approach is anchored by a cutting-edge DG technique, known as MASF [52]. This methodology is built on the foundational principles of MAML [62].

By harnessing the power of MASF, I am able to engage in a fascinating exploration of the extent to which each magnification level can be generalized based on information gleaned from all other magnification levels. Such an innovative strategy offers the tantalizing prospect of creating more streamlined and compact representations for WSI.

In essence, I am investigating how to "translate" information between these different "worlds" of magnification levels, with the ultimate goal of constructing a comprehensive WSI representation that encapsulates the distinctive visual features inherent in each magnification. This pioneering endeavor to illuminate the path toward more effective generalization across magnification levels underpins the essence of my study.

3.2 Magnification Generalization

In the process of implementing the MASF method, let us consider the presence of K training source magnifications, represented by $\{\mathcal{M}_k\}_{k=1}^{K}$. This collection of training source magnifications is integral to my methodology.

To propel the MASF method forward, a specific procedure is followed during each iteration. Primarily, the source magnifications are bifurcated into two distinct categories: meta-train magnifications, denoted by $\mathcal{M}_{\text{tr}}$, and meta-test magnifications, represented by $\mathcal{M}_{\text{te}}$. This division forms the basis of my iterative procedure and paves the way for the application of different loss functions.

In an effort to learn the most suitable latent space representation, I harness the combined power of three distinct loss functions within the MASF framework. The choice of these loss functions is not arbitrary; rather, each of them plays a crucial role in refining the learning process and enhancing the final outcome.

Let us delve a bit deeper into the specifics of these loss functions, as understanding their individual roles and the collective impact they impart on the model is crucial for appreciating the finer nuances of my study. The following sections will offer a detailed overview of these loss functions and their contributions to my work.

3.2.1 Loss for Supervised Learning

The first loss function in my study is the cross-entropy loss for supervised embedding. This loss function is defined as follows:

$$\mathcal{L}_{\text{cr}}(\mathcal{M}_{\text{tr}}; \psi, \theta) := \\ \frac{-1}{|\mathcal{M}_{\text{tr}}|} \sum_{\mathcal{M} \in \mathcal{M}_{\text{tr}}} \frac{1}{|\mathcal{M}|} \sum_{(x,y) \in \mathcal{M}} \sum_{c=1}^{C} \mathbb{I}(y = c) \log \mathbb{I}(\widehat{y} = c), \tag{3.3}$$

Let us break this equation down to understand it more clearly:

1. $\mathcal{M}_{\text{tr}}$ refers to the meta-training magnifications.
2. ψ and θ are the model parameters before updating.
3. C stands for the total number of classes.
4. The term (x, y) represents a tuple where x is a data point in $\mathcal{M}$ and y is its ground truth label.
5. $\widehat{y}$ is the predicted label for the data point x.
6. $\mathbb{I}(\cdot)$ represents an indicator function that outputs 1 when the condition in parentheses is satisfied, and 0 otherwise.

The cross-entropy loss essentially measures the dissimilarity between the ground-truth labels and the predictions made by the model. By averaging the cross-entropy loss across

all patches and classes in the meta-training set, one can quantify how well the model is performing on the task of supervised embedding.

Please note that this loss function is computed in batches in practical scenarios to accommodate the constraints of computational resources.

After updating the model according to the loss of cross-entropy, the new parameters ψ' and ϕ' are obtained. These updated parameters will then be used in the subsequent iterations of the learning process.

3.2.2 Loss for Magnification Generalization

The second loss function in the model is designed to manage the magnification generalization. This is essentially the loss of generalization. Let us delve into the specifics of this loss function:

For a specific magnification level k, consider the embeddings of samples obtained through feature extraction. If samples belonging to the class label c at this magnification are represented as $\{x_{c,i}^{(k)}\}_{i=1}^{n_c}$, their corresponding feature extraction embeddings would be $\{F_{\psi'}(x_{c,i}^{(k)})\}_{i=1}^{n_c}$. With these, one can compute a Monte Carlo approximation for the mean latent space representation of class c at magnification k:

$$z_c^{(k)} := \frac{1}{n_c} \sum_{i=1}^{n_c} F_{\psi'}(x_{c,i}^{(k)}). \tag{3.4}$$

The mean latent space representation, $z_c^{(k)}$, is then processed through the supervised embedding layers, and the softmax function is applied, scaled by a temperature factor $\tau > 1$:

$$s_c^{(k)} := \text{softmax}(F_{\theta'}(z_c^{(k)})/\tau), \tag{3.5}$$

This gives us what can be conceptualized as a "soft" confusion matrix.

Next, for any pair of magnifications $\mathcal{M}_i$ and $\mathcal{M}_j$, one can compute a global loss that is essentially the KL divergence, averaged across all C classes:

$$\begin{aligned} &\ell_{\text{gen}}(\mathcal{M}_i, \mathcal{M}_j; \psi', \theta') \\ &\quad := \frac{1}{C} \sum_{c=1}^{C} \frac{1}{2} \Big[D_{\text{KL}}(s_c^{(i)} \| s_c^{(j)}) + D_{\text{KL}}(s_c^{(j)} \| s_c^{(i)}) \Big], \end{aligned} \tag{3.6}$$

Here, D_{KL} denotes the KL divergence.

This generalization loss is then computed across all the meta-training and meta-testing magnifications:

$$\mathcal{L}_{\mathrm{gen}}(\mathcal{M}_{\mathrm{tr}}, \mathcal{M}_{\mathrm{te}}; \psi', \theta') := \frac{1}{|\mathcal{M}_{\mathrm{tr}}||\mathcal{M}_{\mathrm{te}}|} \sum_{\mathcal{M}_i \in \mathcal{M}_{\mathrm{tr}}} \sum_{\mathcal{M}_j \in \mathcal{M}_{\mathrm{te}}} \ell_{\mathrm{gen}}(\mathcal{M}_i, \mathcal{M}_j; \psi', \theta'), \tag{3.7}$$

The symbol $|\cdot|$ here represents the cardinality, or size, of a set. As with my previous loss function, the computations involved in this equation are performed in batches in practical applications. These computations are represented by Eqs. (3.4, 6.3, and 3.7).

3.2.3 Loss for Metric Learning

Finally, a triplet loss [193] is used for the sake of metric learning. Consider a triplet of anchor, positive, and negative instances denoted by x_a, x_p, and x_n, respectively. Triplet loss attempts to increase the interclass variances of data and decrease the intraclass variances. If R triplets are randomly sampled, $\mathcal{T} := \{x_a^r, x_p^r, x_n^r\}_{r=1}^R$, from all the source magnifications $\{\mathcal{M}_k\}_{k=1}^K$, the average triplet loss is

$$\begin{aligned}\mathcal{L}_{\mathrm{tri}}(\mathcal{T}; \psi', \phi') := \\ \frac{1}{R} \sum_{r=1}^{R} \Big[&\|F_{\phi'}(F_{\psi'}(x_a^r)) - F_{\phi'}(F_{\psi'}(x_p^r))\|_2^2 \\ &+ \|F_{\phi'}(F_{\psi'}(x_a^r)) - F_{\phi'}(F_{\psi'}(x_n^r))\|_2^2 + \zeta \Big]_+,\end{aligned} \tag{3.8}$$

where $\|\cdot\|_2$ is the ℓ_2 norm, ζ is a margin and $[\cdot]_+ := \max(\cdot, 0)$ is the standard Hinge loss.

3.2.4 Updating Parameters

In this section, I delve into the specifics of how the weights of subnetworks, as visualized in Fig. 3.2, undergo iterative updates. The fundamental step of this process involves the use of gradient descent.

To illustrate, let us look at the first step of gradient descent employed to update the weights ψ and θ:

$$(\psi', \theta') := (\psi, \theta) - \alpha \nabla_{\psi,\theta} \mathcal{L}_{\mathrm{cr}}(\mathcal{M}_{\mathrm{tr}}; \psi, \theta), \tag{3.9}$$

In this equation, α represents the learning rate, while $\nabla_{\psi,\theta}$ signifies the gradient with respect to ψ and θ.

To construct the meta loss, I utilize a linear combination of the generalization and triplet losses:

$$\mathcal{L}_{\text{meta}}(\mathcal{M}_{\text{tr}}, \mathcal{M}_{\text{te}}, \mathcal{T}; \psi', \theta', \phi') := \\ \beta_1 \mathcal{L}_{\text{gen}}(\mathcal{M}_{\text{tr}}, \mathcal{M}_{\text{te}}; \psi', \theta') + \beta_2 \mathcal{L}_{\text{tri}}(\mathcal{T}; \psi', \phi'), \tag{3.10}$$

Here, $\beta_1, \beta_2 > 0$ serve as weight coefficients for the respective losses.

Following the first step of gradient descent as defined in Eq. 5.4, the weights are further updated through two additional gradient descent steps:

$$(\psi, \theta) := (\psi, \theta) - \eta \nabla_{\psi,\theta}(\mathcal{L}_{\text{cr}} + \mathcal{L}_{\text{meta}}), \tag{3.11}$$
$$\phi := \phi - \gamma \nabla_{\phi} \mathcal{L}_{\text{tri}}(\mathcal{T}; \psi', \phi'), \tag{3.12}$$

In these equations, η and γ act as the learning rates.

We carry out the above steps of gradient descent iteratively until convergence. The mathematical procedures outlined in Eqs. (5.4, 5.6), and 3.12) are implemented using the backpropagation algorithm. A visual representation of these loss functions is provided in Fig. 3.2. This process ensures that the weights are updated efficiently, driving the learning of my model.

3.3 Experimental Results

In this section, I discuss the experimental setup, including dataset selection, image preprocessing, configuration settings, and results.

3.3.1 Dataset, Preprocessing, and Setup

For the experimental evaluation, the BreaKHis breast cancer histopathology image dataset is employed [214]. Composed of patches derived from four types of benign breast tumors (adenosis (A), fibroadenoma (F), phyllodes tumor (PT), and tubular adenoma (TA)) and four malignant breast cancer tumors (ductal carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC), and papillary carcinoma (PC)), this dataset provided us with a diverse range of samples for investigation. These patches are available in four distinct magnification levels, namely, 40×, 100×, 200×, and 400× according to Fig. 3.3.

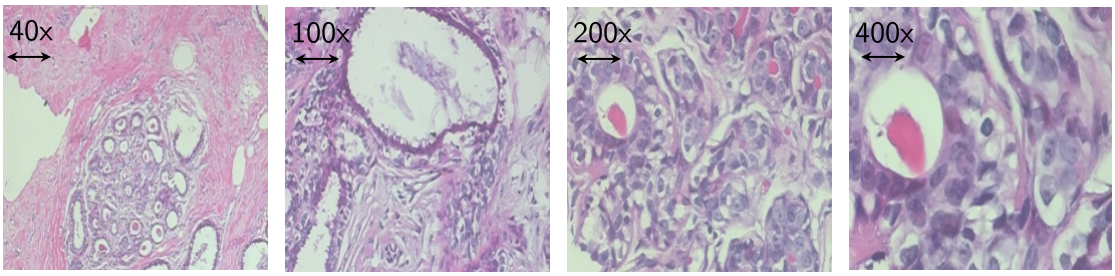


Figure 3.3: Random images of different magnification levels for Adenosis on BreakHis dataset.

To maintain consistency in color and staining across the dataset, the Reinhard stain normalization technique [178] was applied to all images. Then the normalized dataset was split into training, validation, and test sets in a ratio of 45%-45%-10%.

For the backbone, AlexNet was employed [121], with initial weights seeded from pre-trained ImageNet weights [187].

Regarding the hyperparameter settings for the experiments, I adopted the values $\beta_1 = 1, \beta_2 = 0.005, \alpha = \eta = \gamma = 10^{-5}$, which were inspired by the work presented in [52]. For the remaining network configuration and hyperparameters, I kept the setup consistent with the one used in [52]. To optimize the model, I used the Adam optimizer and implemented early stopping as a measure to prevent overfitting.

3.3.2 Magnification Generalization for Tumor Types

Inspired by literature [52], the baseline for comparison of generalization is the ERM approach in which all source domains are combined to train an latent space representation with a cross-entropy loss function. Table 3.1 reports the average accuracy of embeddings by the proposed and ERM models where the accuracies are averaged over three independent runs, inspired by [52]. In this experiment, all eight tumor types are used as classes which is a demanding task due to the presence of similarity of patterns between different tumor types. Four different cases are reported where one domain is left out of the magnification levels to be considered as the target magnification. This generalization is useful in practice for cancer diagnosis of novel magnifications. The table shows my proposed method outperforms the baseline in all cases of target magnifications.

3.3.3 Magnification Generalization for Malignancy Classification

I further put the proposed method to the test by examining its effectiveness in magnification generalization for binary classification. The aim was to classify images based on their malignancy status - that is, whether they were benign or malignant.

Table 3.1: Comparison of magnification generalization for tumor types by ERM and my proposed approach. The rates are accuracy percent.

Source	Target	Magnification Generalization	ERM
100×, 200×, 400×	40×	44.37 ± 0.11	40.39 ± 0.46
40×, 200×, 400×	100×	50.82 ± 0.06	48.82 ± 0.39
40×, 100×, 400×	200×	52.82 ± 0.22	51.75 ± 0.26
40×, 100×, 200×	400×	48.27 ± 0.43	47.21 ± 0.19

Table 3.2: Comparison of magnification generalization for malignancy using my method based on MASF and the ERM method. Accuracy rates are represented as percentages.

Source Magnification	Target Magnification	Magnification Generalization	ERM
100×, 200×, 400×	40×	81.01 ± 0.42	78.86 ± 0.23
40×, 200×, 400×	100×	80.98 ± 0.70	80.52 ± 0.34
40×, 100×, 400×	200×	80.99 ± 0.57	79.96 ± 0.86
40×, 100×, 200×	400×	77.25 ± 0.83	74.44 ± 0.31

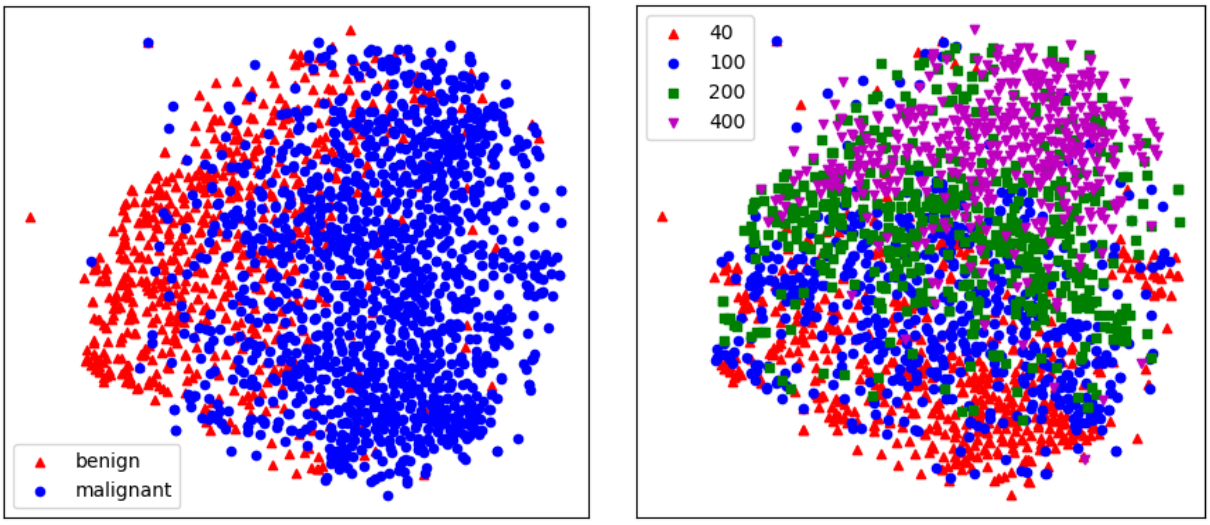
For this task, I compared the performance of the method against the ERM approach. The mean accuracy rates of both methods were calculated over three independent runs and presented in Table 3.2. Given that binary classification is generally simpler than multi-class categorization, the accuracy rates reported here are higher than those of previous experiments. It should be noted, however, that the rates are slightly lower than those achieved with magnification adaptation methods, such as those presented in [12, 42, 215]. This is primarily due to the fact that generalization is a more challenging task than adaptation, as indicated by [52].

As indicated in Table 3.2, my proposed method outperformed the ERM baseline across all target magnifications, which speaks to its effectiveness.

Figure 3.4 provides a visualization of a trained embedding example, with a target magnification of 400×. Here, the embedding is plotted with labels indicating both malignancy and magnification. The figure demonstrates that classes, including those with an unseen target magnification, are well separated. It is also evident that both benign and malignant instances include a range of different magnification levels, including the 400× level, which was not seen during training. The coherence of the trained latent space representation further underscores the effectiveness of the presented magnification generalization method.

3.4 Conclusion and Summary

In this chapter, I introduced a method for generalizing different magnification levels in histopathology images. The unique feature of my method is that during the training phase, the target magnification level is kept unknown to the model. The foundation of my proposed model is the MASF framework, which itself was built on the concept of MAML,



(a) (b)

Figure 3.4: latent space representation visualization of the test dataset with the target magnification set at 400×. (a) malignancy-wise, (b) magnification-wise

specifically designed for DG.

The experiments utilized the BreaKHis dataset and the results obtained confirmed the robustness and effectiveness of the approach over the ERM. The introduced method demonstrated a significant improvement in the accuracy of classifications, especially when the target magnification level was unknown during the training phase.

In the following chapters, I will be detailing the studies I have conducted concerning the OOD generalization applicable to unseen hospital settings.

Chapter 4

Generalization of Vision Pre-trained Models to Unseen Hospitals

Prologue

The content of this chapter is based on an article published during my Ph.D. research: Generalization of vision pre-trained models for histopathology- M Sikaroudi, M Hosseini, R Gonzalez, S Rahnamayan, HR Tizhoosh, Nature, Scientific Reports 2023 [207]

4.1 Introduction

Artificial neural networks offer the capability to fit model weights to data, thereby yielding highly precise outputs. However, difficulties arise when attempting to generalize these models to unseen data. Various terminologies have been used in the literature to describe these challenges. Some studies suggest that poor OOD generalization is a result of the models learning "shortcuts" [67, 148, 182] or "biases" [232, 203, 44].

Contrarily, other research approaches the issue of OOD generalization from a unique angle. They assert that the main cause of subpar OOD performance is the "domain shift" that occurs between the source and target domains [216, 149, 217].

Different nomenclatures can be summarized as follows:

- **Bias**

  Definition: Bias is an inherent or acquired inclination towards favoritism or prejudice that is targeted towards a specific entity or a group of entities. This often results in unfair treatment or judgment. Bias, in this context, can significantly influence the decision-making process in AI models, often leading to skewed or discriminatory outcomes [156].

Illustration: A clear example of bias in AI can be seen in the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system. COMPAS is a decision-making tool used by judges to assess an offender's potential risk of recidivism - the likelihood of a convicted criminal to re-offend. It plays a crucial role in determining if an offender should be released or retained in prison. However, an investigation uncovered that the COMPAS algorithm had an inherent bias against African-Americans, treating them more unfavorably compared to other groups [156].

Related Studies: The issue of bias in AI and machine learning has been a significant focus of research. Various studies explore the concept and its implications, offering insight into how bias can affect models and their decision-making processes [232, 203, 44].

Consequence: The inherent bias in AI models contributes to the challenge of generalizing these models to unseen data. In essence, the models may not perform accurately or fairly when presented with new data that falls outside of their training domain. This lack of generalization raises concerns about the reliability and fairness of AI models.

- **Shortcuts**

  Definition: The term "shortcut" refers to a decision rule that works effectively on i.i.d. test data but does not perform as expected when applied to OOD test data. This situation results in a discrepancy between what the AI model is designed to learn and what it actually learns [67].

  Illustration: An example of an AI system taking a "shortcut" can be observed in image recognition models. For instance, when such a model trained on cows in a grassland context is presented with an image of a cow on a beach, it may fail to correctly classify the cow. This happens because the model may use the usual background as a significant feature for recognizing cows. Therefore, when the context shifts to an unexpected environment (the beach), the model may incorrectly classify the image [14].

  Related Studies: Numerous studies have delved into the issue of "shortcuts" in AI models. They have highlighted the challenges and potential solutions in understanding and rectifying the issues caused by these "shortcuts" [67, 148, 182].

  Consequence: The adoption of shortcuts by AI models contributes to the difficulty in generalizing these models to unseen or new data. In other words, these models might not perform accurately when they encounter data that significantly differs from the data they were trained on. This can cause serious issues in real-world applications where data often come from varied and unpredictable sources.

- **Domain/Distribution Shift**

  Definition: Domain or distribution shift refers to the discrepancies or changes between the source data (on which a model is trained) and the target data (on which

the model is tested or deployed). This concept is particularly significant in the field of transfer learning, where models are trained on one type of data and then applied to another [120].

Illustration: Various factors can contribute to a domain shift in visual data. For instance, sampling bias can lead to differences in the types of images that are collected. Additionally, variations in image content or perspective can create discrepancies. Even changes in image characteristics such as brightness, noise level, or color can lead to significant domain shifts. These shifts can greatly affect the performance of machine learning models when they are applied to new data that differ from the training data in these respects [120].

Related Studies: There are numerous studies addressing the concept of domain shift and its effects on model performance. These works provide insights into how domain shifts occur and how they can be managed or mitigated to improve model generalization [149, 217, 216].

Consequence: The issue of domain shift plays a critical role in inducing challenges related to model generalization on unseen data. When a model encounters data that differs significantly from its training data, its performance may degrade, leading to unreliable or inaccurate outcomes. This raises concerns about the robustness of AI models in dealing with real-world, diverse data.

In the field of histopathology, Hägele et al. [80] have defined three distinct categories of biases that occur in histopathology setups. Though they have provided specific terminologies, these types of biases bear strong resemblances to the concept of "shortcuts" in the broader machine learning literature. These biases can lead to satisfactory in-distribution performance but poor OOD performance. Here is an elaboration of these categories:

- **Dataset Bias**
  This form of bias occurs when only a small portion of an image is correlated with its class label. For instance, in a situation where a small central area of each image signifies the class label and the remaining areas are irrelevant, a deep learning network may struggle to generalize. The issue arises when the model is tested on images where the subject of interest does not necessarily occupy the center. This can lead to errors in prediction and hamper the model's ability to generalize effectively to new data.

- **Label Bias**
  Label bias refers to biases that accidentally correlate with class labels. Suppose a deep learning network is trained on a set of images where a unique red spot characterizes a particular class. If the model generalizes this feature as a characteristic of that class, it may fail to recognize the same class in the test images that lack this unique red spot. Such a bias can limit the model's ability to effectively generalize to new data.

- **Sampling Bias**
  Sampling bias happens when the training dataset lacks certain critical features or textures. For instance, if a network is trained on images that do not include certain tissue textures such as necrosis, its performance may degrade when tested on images with these previously unseen textures. This is because the network has not *seen* or *learned* these textures during the training phase.

In addition to defining these bias categories, Hägele et al. also demonstrated the value of Explainable Artificial Intelligence (XAI) techniques for visualizing and understanding these biases [80]. By doing so, they contribute to a better understanding of how biases in data can affect model performance, and how these biases can be identified and mitigated to improve OOD generalization.

To effectively deploy deep learning models in real-world environments, it is of paramount importance to account for potential distribution shifts that may occur between source and target data. This need is particularly pronounced in fields such as digital pathology, where variations in data acquisition methods across trial sites or over time may introduce domain shifts. Such shifts can result from subtle and possibly visually unnoticeable differences among WSIs, significantly impacting model performance and reliability.

A comprehensive understanding of distribution shift, its causes, and its implications is crucial to fully exploit the vast potential that deep learning promises in the field of histopathology. It is vital to ensure that a model's predictions remain trustworthy and accurate, even when new, unseen data is introduced. Although this task of correctly modeling and responding to unencountered data during training is undoubtedly challenging, several methods have recently been proposed to improve OOD generalization.

Among these approaches, multi-domain learning regimes, such as domain generalization and domain adaptation, have shown promise. These strategies aim to enhance the model's performance on OOD data by employing specialized training methods. They primarily fall into three categories:

1. **Simulating OOD Data During Training**: This approach generates synthetic data that mimic the characteristics of potential OOD data, providing the model with exposure to such data during training [134, 52, 208].

2. **Learning Invariant Representations**: This strategy aims to learn representations that remain constant across various data domains. Such invariant representations can help improve model robustness to domain shifts [4].

3. **Creating Adversarial Data Acquisition Scenarios**: This involves deliberately introducing challenging data acquisition scenarios to the model during training. This method can prepare the model to handle difficult or unexpected situations that might occur in the real world [242].

By exploring and implementing these techniques, we can better equip deep learning models to handle the challenges of distribution shifts, ultimately improving their performance and reliability in real-world, OOD scenarios.

Although domain generalization is a field that has been considerably explored [266], recent studies have questioned the efficacy of existing strategies [76, 192]. For instance, Wiles et al. [250] delved into the effect of three types of distribution shifts: (1) spurious correlations, (2) low-data drifts, and (3) unseen shifts.

Their findings were mixed and did not conclusively favor a particular method. However, they highlighted that rudimentary techniques such as data augmentation and pre-training were often effective. Additionally, they demonstrated that domain generalization algorithms show efficacy in handling certain types of data and distribution shifts. However, the optimal approach varied based on the dataset and the nature of the attribute, underlining the absence of a one-size-fits-all solution and emphasizing the need to enhance algorithmic robustness for diverse real-world scenarios.

These observations led to a critical question: Has domain generalization truly advanced beyond standard ERM algorithms? [250]. Such contemplations might appear discouraging at first; however, they are crucial for the evolution of the field. On a positive note, numerous other studies underscore that machine learning models can indeed be generalized across datasets that exhibit different distributions [250, 176].

For instance, several works have supported the efficacy of pre-training on large datasets for improving OOD generalization [226, 250]. Such strategies can potentially provide the model with a broad understanding of different data characteristics, thus enabling it to cope better with unseen data or distribution shifts.

In essence, while current domain generalization techniques provide some promise, there is a need for further research and exploration to identify more robust and universally effective methods. This ongoing endeavor is crucial to ensure that machine learning models can consistently deliver reliable performance across diverse real-world scenarios.

This chapter delves into a comprehensive exploration of the role and efficacy of pre-trained models in achieving OOD generalization. This investigation includes a wide range of pre-trained models, some of which are trained on natural images while others employ histopathology images. An important aspect of this research is the utilization of a leave-one-*hospital*-out cross-validation method. This method involves isolating each WSI repository associated with individual hospitals and subsequently fine-tuning the pre-trained models using the remaining WSI repositories for the task at hand.

The primary goal of this chapter is not simply to achieve top-tier results on benchmark datasets but rather to enhance our understanding of how pre-trained models can contribute to more robust OOD generalization. By shedding light on the mechanisms and effectiveness of pre-trained models in managing OOD generalization, we aim to generate valuable insights that could help narrow the gap between in-distribution and OOD performance in future research endeavors.

The research contribution of this chapter can be summarized into three primary areas:

1. **Significance and Nature of Pre-training for OOD Generalization**: It was demonstrated that pre-training on large datasets plays a vital role in OOD generalization.

This observation holds true for both Semi-WSL [258] and Semi-SL [258] as compared to a basic ImageNet [45] pre-trained model. However, the specific nature of the pre-trained model also significantly influences OOD generalization. For instance, KimiaNet [180] exhibits distinct performance when compared to Semi-WSL [258] and Semi-SL [258]. The experimental results indicate that lacking either of these components - size of the pre-training dataset or appropriate nature of the pre-trained model - may lead to a decline in OOD generalization.

2. **Role of Fixed-Policy Augmentations**: The experiments suggest that with fixed-policy augmentations, OOD generalization can be improved by reducing reliance on shortcuts and instead focusing more on semantically interpretable features. However, this approach comes with its challenges, as it could complicate the training of deep networks. In essence, fixed-policy augmentation can be both beneficial and detrimental, depending on the nature of the OOD test data. As such, one cannot assume a one-size-fits-all fixed-policy augmentation that will work across all scenarios.

3. **Correlation between In-Distribution Performance and OOD Performance**: This study shows that improving in-distribution performance does not always translate into better OOD performance. This finding challenges the commonly held belief that in-distribution performance is a reliable indicator of OOD performance [2], thereby highlighting the need for specific metrics or methods to assess and ensure OOD generalization.

By illuminating these aspects, it is hoped that future research endeavors will be guided toward the development of more robust and universally effective methods for achieving high-quality OOD generalization across various application scenarios.

In these experiments, a variety of pre-trained models were employed. These include (1) the vanilla ImageNet model, (2) the Semi-SL model trained on both ImageNet and YFCC100M, (3) the Semi-WSL model trained on ImageNet and 940 million public images tagged with 1.5K hashtags, and (4) KimiaNet, which is trained on TCGA dataset. The summary of these pre-trained models and their details can be found at Table 4.1.

Pre-trained Model	Architecture	Number of Parameters	Pre-training Data	Feature Space Dimension
Vanilla	ResNet18	11,689,512	ImageNet	512
Semi-SL	ResNet18	11,689,512	YFCC100M, ImageNet	512
Semi-WSL	ResNet18	11,689,512	IG-1B-Targeted with 1.5K hashtags, ImageNet	512
KimiaNet	DenseNet121	7,978,856	Subtying of TCGA WSIs	1024

Table 4.1: Details of pre-trained models used in the study.

4.2 Experimental Setup and Methods

When investigating OOD performance in histopathology configurations, most researchers typically rely on datasets derived from TCGA [145, 131, 208]. However, the KimiaNet model [180], which forms a key part of my research, has been pre-trained on all WSIs from

the TCGA dataset. Consequently, using TCGA data to define OOD test set would not be appropriate, given that it has already been leveraged during the pre-training phase of the model. This constraint prompts us to look for alternative data sources to construct the test set.

Among available options, the CAMELYON17 dataset emerges as a suitable choice due to its multi-hospital representation. The dataset's heterogeneity makes it well-suited for assessing the OOD performance of the pre-trained models. In the subsequent sections, I provide a detailed description of the data and models utilized in the study and outline the structure of the experimental setup. These detailed outlines are designed to offer insights into the research methodology and enable replication and extension of the study.

4.2.1 The CAMELYON17 Dataset

The CAMELYON17 dataset [9], a rich source of histopathological data, comprises 1000 WSIs obtained from five distinct medical centers. These WSIs present not only disparate variations in stain colors [229] but also differences in morphology and tumor staging across the participating medical centers [8, 138] (as illustrated in Fig. 4.1).

In the CAMELYON17 challenge, 500 WSIs were allocated for training, and the remaining 500 were utilized for testing purposes. The training subset of CAMELYON17 includes 318 WSIs categorized as negative and 182 WSIs identified as having metastases. However, out of all the slides, only 50 WSIs included pixel-level annotations. Thus, for the purpose of sampling tumor and non-tumor cells, only these 50 slides were considered. Although sampling non-tumor cells from the other slides could introduce further variations, the impact on the overall results is not expected to be substantial [217].

It is also important to note that tumor areas typically occupy only a minor portion of the slide area. This disparity leads to a significant patch-level imbalance. To counteract this issue, a patch sampling strategy mirroring the approach outlined in [139] was implemented. In this strategy, an equal number of tumor and normal patches are sampled from each slide, ensuring a uniform distribution of patches.

As a result of this sampling strategy, approximately 3000 patches were acquired from each hospital. These patches comprised a balanced distribution of tumor and non-tumor samples, with each category constituting about half of the total patches.

4.2.2 Defining the OOD Hospital and Data Segments

In this research, the "leave-one-hospital-out" approach is employed. Here, for each hospital, designated as H_{external}, the model's backbone is trained exclusively on images derived from the other hospitals, collectively termed H_{internal}.

The H_{internal} data is further segmented into distinct chunks for training, validation, and in-distribution testing. These segments account for 70%, 10%, and 20% of the H_{internal} data, respectively.

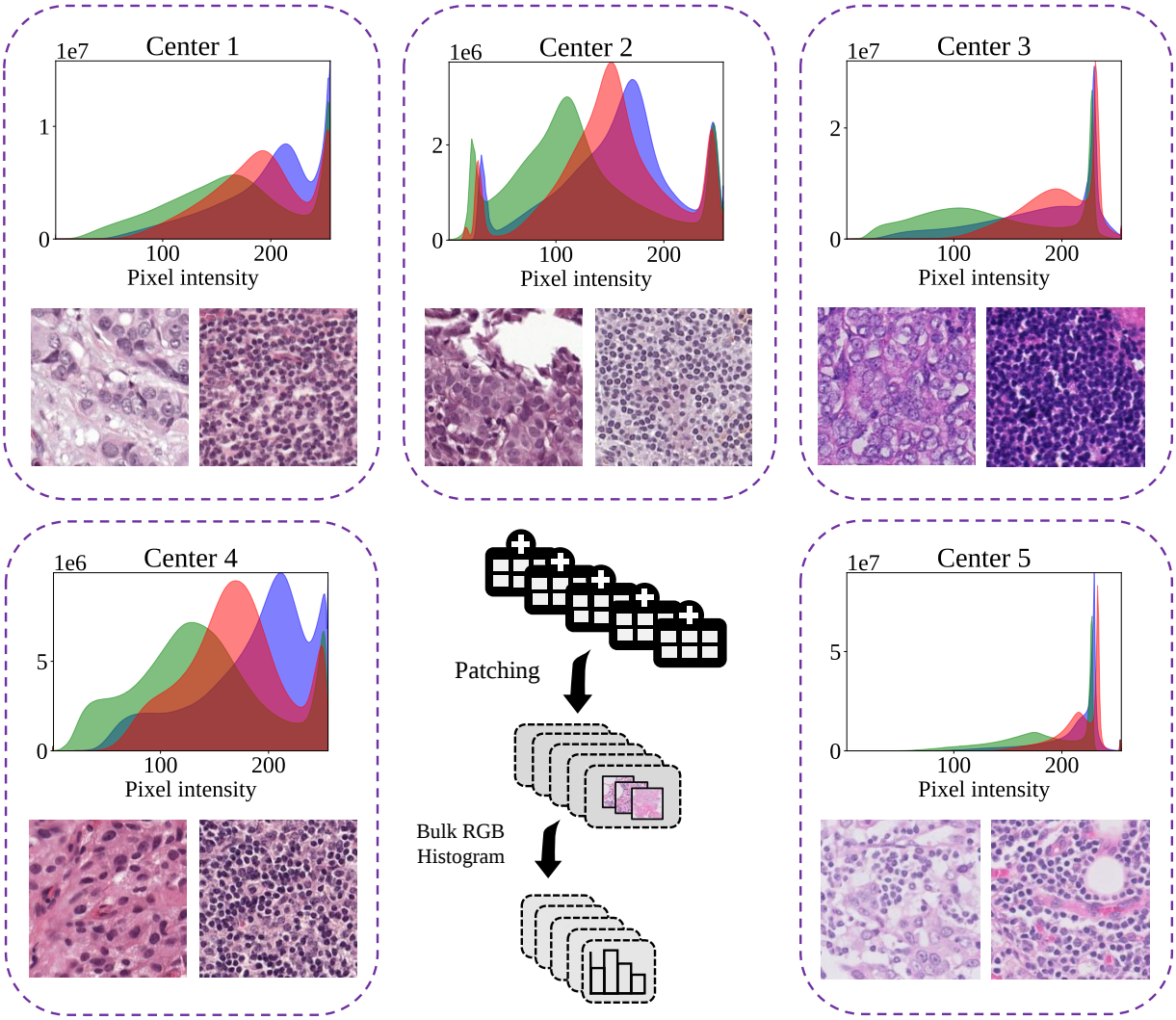


Figure 4.1: The bulk RGB histogram of the 512×512 extracted patches as well as sample tumor and non-tumor patches of each center/hospital in the CAMELYON17 dataset. Hospitals 3 and 5 have quite different histograms in comparison to the rest of the hospitals.

As the training progresses through each epoch, the accuracy is computed for both the external and in-distribution datasets. These calculations yield the OOD top-1 accuracy and in-distribution top-1 accuracy, which serve as performance indicators for the model's ability to generalize across various hospitals.

4.2.3 Variations in Training Data Scenarios

This section introduces various scenarios that are put forward for the fine-tuning or training of the models involved in the experiments. For this purpose, three distinct scenarios for the

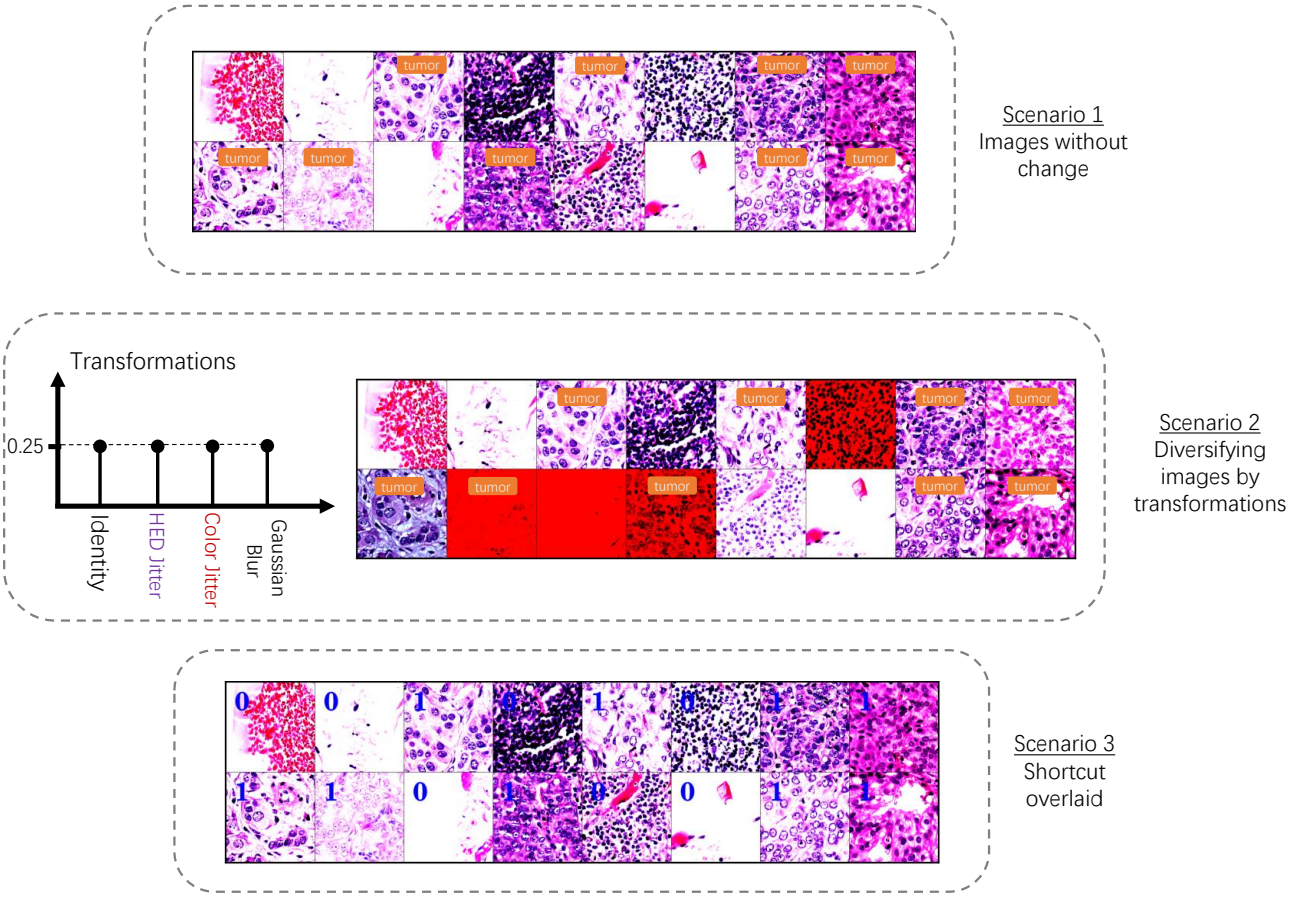


Figure 4.2: An example training batch for different scenarios. It is noteworthy that in Scenario 1, the training set patches are devoid of any form of augmentation. As can be seen in the figure, in Scenario 2, one transformation is selected from the set of *identity*, *HED jitter*, *color jitter*, and *Gaussian blurring* transformations (with uniform distribution ($p = 0.25$)) for each image in the batch. In Scenario 3, the correct label (0: non-tumor, 1: tumor) for each image is overlaid on the image itself.

training are evaluated, as illustrated in Fig. 4.2. The delineated scenarios are as follows:

Scenario 1: In this scenario, the training images, identified as H_{internal}, are introduced into the network without undergoing any alterations.

Scenario 2: This scenario presents a case where multiple types of distortions, common in histopathology setups (refer to Fig. 4.2), are simulated and uniformly applied at random to H_{internal}, prior to feeding these images to the deep learning network. The transformations considered are as follows:

- HED jitter [228]: This method introduces random perturbations to the HED color space value of an RGB histopathology image. The hematoxylin and eosin color channels are initially separated using a color deconvolution technique [186]. The

hematoxylin, eosin, and Diaminobenzidine (DAB) stains are then independently perturbed. The resulting stains are subsequently converted back to the standard RGB color space. These manipulations facilitate stain invariance in the model.

- Color jitter: This transformation directly alters image attributes such as brightness, contrast, and saturation to enhance image diversity, thereby boosting the model's robustness against color variations.
- Gaussian blurring: This method applies a blurring effect using a Gaussian kernel with a defined radius.
- Identity: This transformation leaves the input images unchanged.

Scenario 3: This scenario involves overlaying a digit (0: non-tumor and 1: tumor) corresponding to the label of images on the top left corner of each image, as illustrated in Fig. 4.2. This experiment is set aside for subsequent sections where the concept of shortcut learning is discussed.

4.2.4 Implementation of Pre-trained Models in Training

Several pre-trained models have been leveraged and evaluated for this study, including vanilla ImageNet, Semi-SL [258], and Semi-WSL [258] pre-trained models. All these pre-trained models are constructed on the ResNet18 backbone [84].

In addition to these pre-trained models on natural images, an investigation has also been conducted into the performance of the KimiaNet pre-trained model [180]. Unlike the previous models, KimiaNet is domain-specific, having been trained explicitly for histopathology applications. This provides an opportunity to evaluate the performance of a model that has been primed for the specific demands and peculiarities of histopathology imagery.

The implementation of these models in the experiments followed certain established guidelines and settings. Consistent with the studies of [136, 261], the total batch size was fixed at 32. For the learning rate, a value of 0.01 was adopted for the instances where training was conducted from scratch. Conversely, for pre-training instances, the learning rate was reduced to 0.001. This was combined with a step-LR schedule comprising of 7 steps and a γ value of 0.1.

The choice of optimizer is a critical component in the training process. For these experiments, the SGD optimizer was utilized. The SGD optimizer has seen widespread use in the literature [184] due to its efficiency and simplicity. To regularize the model and prevent overfitting, a weight decay value of $1e-4$ was used.

4.3 Analysis and Interpretation of Results

Throughout the course of the conducted experiments, certain notable trends and patterns emerged. One of the most striking observations pertains to the differential progress of the in-distribution test accuracy and the OOD top-1 accuracy throughout the training process.

While the in-distribution test accuracy displayed a near-constant upward trend, the OOD top-1 accuracy was not as consistent. This discrepancy may hint towards an intriguing facet of the learning process taking place during the training. It suggests that the models do not only learn semantic features that align with the general characteristics of the class, but they also pick up on non-semantic or hospital-specific features. These non-semantic features, rather than enhancing the model's generalization ability, could paradoxically be undermining it. This potential undermining is evidenced by the unstable progress of the OOD top-1 accuracy, as these hospital-specific features may not be applicable or helpful when the model encounters data from a different hospital (or distribution).

In the ensuing sections, the focus will shift toward a comparative assessment of different types of pre-trained models. This comparison aims to discern and elucidate the differences in their out-of-distribution performance, providing valuable insights into the strengths and limitations of each model when it comes to handling OOD data. This analysis can offer a foundation for future research efforts aimed at improving OOD generalization and could guide the choice of models and training strategies in practical histopathology applications.

4.3.1 Out-of-Distribution Performance of Pre-Trained Models

Comparison between Training from Scratch and Using Pre-Trained Models: One of the most salient findings drawn from the conducted experiments is the considerable advantage that pre-trained models exhibit over models trained from scratch when it comes to out-of-distribution generalization.

The data detailed in Table 4.2 elucidate the substantial gap in performance between these two approaches. On average, pre-trained models yield significantly superior results, underscoring the merit of leveraging prior knowledge in tackling the challenge of out-of-distribution generalization. Consequently, it can be inferred that employing any reasonably pre-trained model tends to be more beneficial than resorting to training from scratch, particularly when the objective involves handling OOD data.

These observations align with the findings reported in previous research, such as the study by Yu et al. [261]. This consensus reinforces the argument in favor of pre-trained models and advocates for their adoption in applications where out-of-distribution generalization is of paramount importance.

It is crucial to note, however, that the choice of pre-training depends on the specific task at hand and the availability of relevant pre-trained models. Future studies might delve into examining the impact of varying degrees of relatedness between pre-training and target tasks, and how that affects out-of-distribution performance.

When it comes to the *training-from-scratch* approach, Table 4.2 reveals that *scenario 2* typically falls short of the performance achieved by *scenario 1*. The situation seems to take a turn when pre-trained models enter the picture. As per the data in Table 4.2, *scenario 2* demonstrates superior performance to *scenario 1* when hospitals 2, 3, or 5 are used as the hold-out dataset.

This discrepancy suggests an intriguing interplay between the initial weights of the deep learning model and the complexity of the training process. With a suitable foundation of initial weights (achieved through pre-training), the model appears to benefit from additional complexity (in the form of augmentation or diversification) during training, resulting in enhanced generalization to out-of-distribution data. However, if the model does not have the advantage of appropriate initial weights (as in the case of training from scratch), the added complexity seems to hinder the learning process, possibly leading the model astray from capturing meaningful and semantic features.

Comparison between Vanilla, Semi-SL, and Semi-WSL Pre-training: An analysis of the maximum performance on each hold-out hospital, highlighted in Table 4.2, presents some compelling observations. Remarkably, neither the models trained from scratch nor the vanilla pre-trained models gain any highlights.

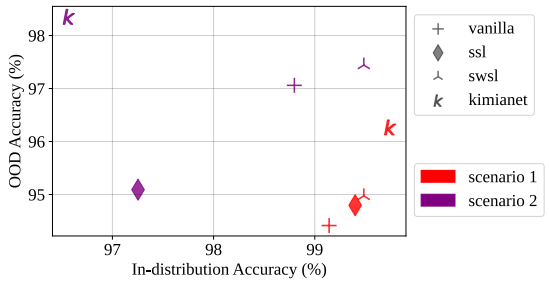
This trend accentuates the potential of Semi-SL [258] and Semi-WSL [258] pre-training strategies as robust alternatives to the conventional vanilla pre-training. This superior performance can likely be ascribed to the extensive and representative datasets on which the Semi-SL and Semi-WSL models have been pre-trained, enabling them to capture and retain more generalized features.

However, there are instances where training under *scenario 2* has resulted in diminished OOD performance. A case in point is when hospital 2 served as the hold-out set. Given the relatively smaller image set from this hospital (roughly 2000 images compared to 3000 from others), it is speculated that image diversification or augmentation may actually impair performance by introducing unwarranted complications to the training process of the deep learning model.

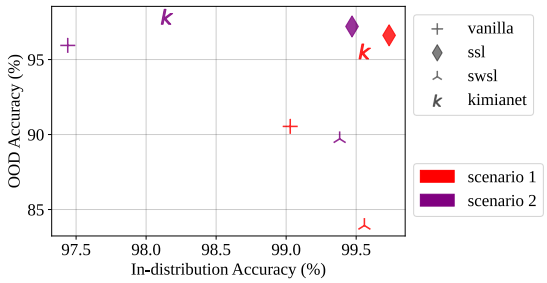
Exploring KimiaNet: The performance of the KimiaNet model has been examined

Table 4.2: The OOD performance of *training from scratch* versus the pre-trained models (vanilla, Semi-SL, and Semi-WSL). Each column represents the OOD top-1 accuracy on the hold-out set.

Pre-training	Weights	Training scenario	**Hospital 1**	**Hospital 2**	**Hospital 3**	**Hospital 4**	**Hospital 5**	*Average*
F	Random	S_1	93.01	89.22	84.95	91.06	81.09	87.9 ± 4.22
F	Random	S_2	92.72	90.28	82.01	90.00	80.71	87.1 ± 4.73
T	Vanilla	S_1	98.75	96.03	94.42	96.65	90.54	95.3 ± 2.69
T	Vanilla	S_2	98.62	93.60	97.06	97.19	91.67	95.6 ± 2.52
T	Semi-SL	S_1	98.52	96.92	94.8	97.46	96.61	96.9 ± 1.19
T	Semi-SL	S_2	99.18	94.98	95.09	97.79	97.21	96.8 ± 1.59
T	Semi-WSL	S_1	99.08	96.52	94.97	98.12	83.93	94.5 ± 5.37
T	Semi-WSL	S_2	99.31	96.19	97.44	98.09	89.71	96.1 ± 3.3
Average			97.4 ± 1.95	94.2 ± 2.05	92.6 ± 4.01	95.8 ± 2.29	88.9 ± 4.48	



Hold-out set: Hospital 3



Hold-out set: Hospital 5

Figure 4.3: The OOD versus in-distribution top-1 accuracy for the model trained using *scenario 1* versus *scenario 2* for the hospitals 3 and 5 with significant distribution shift relative to other hospitals.

under two distinct settings: (1) linear probing [125] (where the feature extractor remains frozen while the classification head alone undergoes training), and (2) fine-tuning (where all model parameters are updated). As summarized in Table 4.3, it is clear that fine-tuning offers superior results compared to linear probing.

Interestingly, the average results across hospitals 2, 3, and 5 appear to be both lower and more variable when compared to hospitals 1 and 4. Moreover, *scenario 2* training surpasses *scenario 1* when the hold-out trial site is either hospital 1, 3, or 5.

Looking at both Table 4.2 and Table 4.3 provides some enlightening insights. KimiaNet, a domain-specific (histopathology) pre-trained model, proves to be superior to all other pre-trained models in at least three of the five external validation cases. This suggests that pre-training models using domain-specific data can significantly enhance out-of-distribution generalization. However, it is important to note that while linear probing outperforms *training from scratch* in both *scenario 1* and *scenario 2*, it falls short of the performance achieved by any of the fine-tuning approaches, regardless of the pre-trained model used.

However, it is important to note that while linear probing outperforms *training from scratch* in both *scenario 1* and *scenario 2*, it falls short of the performance achieved by any of the fine-tuning approaches, regardless of the pre-trained model used.

Analyzing Performance Variations in Hospitals 3 & 5 - OOD versus In-Distribution:

Table 4.3: The OOD performance of linear-probing versus the fine-tuning of KimiaNet. Each column represents the OOD top-1 accuracy on the hold-out (external) hospital.

Training	Scenario	Hospital 1	Hospital 2	Hospital 3	Hospital 4	Hospital 5	*Average*
Fine-tuning	S_1	98.75	96.6	96.44	98.58	95.54	97.2 ± 1.24
	S_2	99.18	95.95	99.18	98.45	97.85	98.1 ± 1.17
Linear-probing	S_1	97.59	89.79	92.45	95.58	80.57	91.2 ± 5.82
	S_2	96.77	86.77	94.71	96.7	91.62	93.3 ± 3.69
Average		98.1 ± 1.08	92.3 ± 4.69	95.7 ± 2.78	97.3 ± 1.42	91.4 ± 7.51	

Tables 4.2 and 4.3 reveal distinct variations in performance accuracy between the pre-trained models and *training-from-scratch*, particularly when hospitals 2, 3, and 5 are set as the hold-out hospitals. Notably, the performance of deep networks is at its lowest among all holdout hospitals under these circumstances.

The lower performance and variability of results when Hospital 2 serves as the hold-out hospital is understandable given its smaller quantity of patches (approximately 2000 versus approximately 3000 in other hospitals). However, the results obtained for hospitals 3 and 5 necessitate further investigation, as their variability could indicate a significant distribution shift in these medical centers compared to others. To visualize this, the OOD versus in-distribution accuracies have been plotted in Figure 4.3.

OOD performance is of crucial importance when considering real-world applications. When examining the different pre-trained models, it becomes evident that KimiaNet outperforms the others in terms of OOD performance when Hospital 5 is the hold-out set. The Semi-WSL [258], vanilla, and Semi-SL [258] pre-trained models follow next in performance when considering training under *scenario 2*. In contrast, under *scenario 1*, KimiaNet retains the leading position with Semi-WSL, Semi-SL, and vanilla pre-trained models ranking subsequently.

Observations on Training Scenarios and Performance:

One crucial observation from the experiments is that training under *scenario 2* consistently improved OOD performance more effectively than *scenario 1*, despite lowering in-distribution performance across all pre-trained model types. In essence, the various transformations utilized in *scenario 2* effectively enhanced OOD performance but had the unintended side effect of degrading in-distribution performance. This finding underscores the fact that in-distribution accuracy does not necessarily correlate with or predict OOD performance. An illustrative example can be found in the performance of KimiaNet under *scenario 2*, which, despite having the least impressive in-distribution performance, delivered the best OOD performance. Additionally, it is worth noting that when Hospital 3 was the hold-out hospital, KimiaNet under *scenario 1* delivered the best in-distribution and OOD performance. However, the adoption of *scenario 2* bolstered the OOD performance, albeit at the expense of in-distribution performance. This intriguing phenomenon can be attributed to *shortcut learning*, which, based on the current understanding, offers satisfactory in-distribution performance while undermining OOD performance. Consequently, this case warrants further investigation using XAI techniques to illuminate potential shortcuts and offer more profound insights into the observed performance dynamics.

This intriguing phenomenon can be attributed to *shortcut learning*, which, based on the current understanding, offers satisfactory in-distribution performance while undermining OOD performance. Consequently, this case warrants further investigation using XAI techniques to illuminate potential shortcuts and offer more profound insights into the observed performance dynamics.

Exploring the Concept of *Shortcut* Learning in Neural Networks:

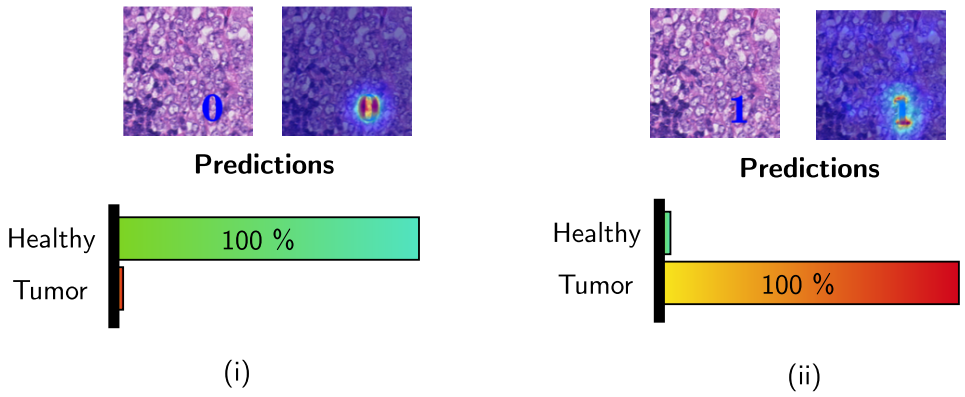


Figure 4.4: KimiaNet trained using *Scenario 3* when tested with a tumorous OOD patch with different class labels overlaid and their corresponding GradCAM heatmaps. **(left)** When false label (0: non-tumor) has been overlaid on the image. According to the class prediction of the network, the network has thoroughly paid attention to the overlaid digit and misclassified the image with its misleading shortcut. **(right)** When the true label (1: tumor) has been overlaid. The network, by focusing on the shortcut, classified the patch with a high degree of certitude.

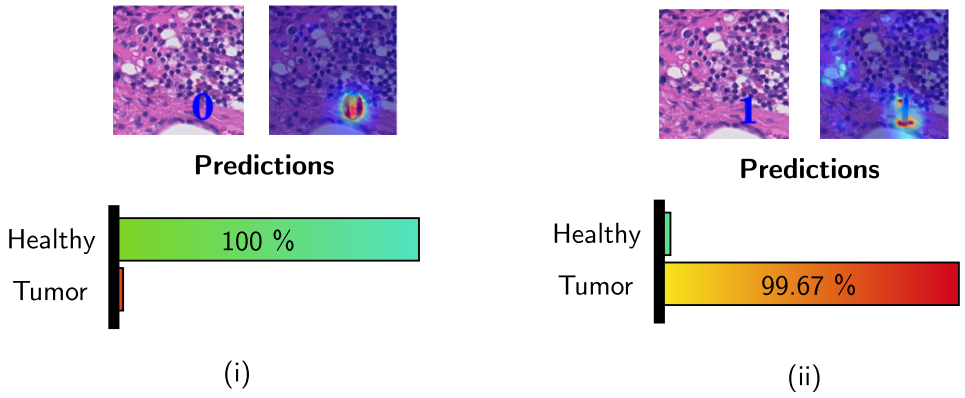


Figure 4.5: KimiaNet trained using *Scenario 3* when tested with a healthy (non-tumor) OOD patch with different class labels overlaid and their corresponding Grad-CAM heatmaps. **(left)** When true label (0: non-tumor) has been overlaid on the image. The network, by relying on the shortcut, classified the patch with confidence. **(right)** When the false label (1: tumor) has been overlaid. According to the class prediction of the network, the network has thoroughly paid attention to the overlaid digit and misclassified the image with its misleading shortcut.

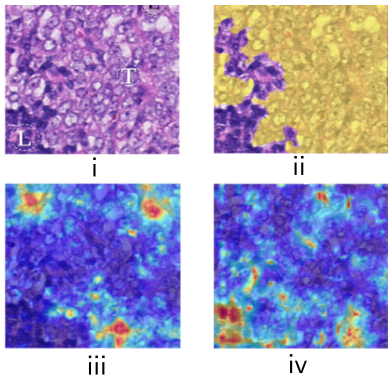


Figure 4.6: The result of training using *scenario 1* and *scenario 2*: **(i)** an OOD tumorous patch (from hospital 3) with different anatomical structures, Ⓣ: Tumor cells, Ⓛ: Lymphocyte, Ⓔ: Erythrocyte. **(ii)** Expert annotation for tumorous regions. **(iii)** GradCAM heatmap for the model trained using *scenario 2* which correctly classified the patch, **(iv)** GradCAM heatmap for the model trained using *scenario 1* which misclassified the patch as a healthy patch.

The general function of neural networks, and machine learning algorithms more broadly, involves establishing decision rules that map inputs to corresponding outputs. For example, in classification tasks, the aim is to assign a specific category to each input image. However, an intriguing phenomenon often observed in these networks is the reliance on what is termed as *shortcuts*. When a network depends on these *shortcuts*, it performs well on the training data and in-distribution tests, yet its performance on OOD tests is usually suboptimal. This discrepancy reflects a disparity between the intended solutions, based on robust general principles, and the solutions the network has actually learned, which often take advantage of idiosyncratic features in the training data that may not generalize well [67].

These shortcuts, while expedient for in-distribution data, may lead to subpar performance when facing novel or more complex data structures, as seen in OOD tests. The exploration of this phenomenon allows us to better understand and address the limitations of current machine learning algorithms, as well as develop more reliable and robust models for diverse and unpredictable real-world scenarios.

Distinguishing Between Shortcut and Bias:

In the realm of machine learning, the term "bias" is used to denote any form of preferential treatment given to a certain entity [86]. This preferential treatment could be directed towards a specific demographic, a certain set of data originating from a particular hospital, or even a subset of data with certain unique characteristics. This favoritism may or may not pave the way for shortcut learning. It can be postulated that a bias is likely to exist if the training of a deep network only includes images from a single, specific trial site. Consequently, a deep network that is skewed or biased might be trained, which may

or may not perform satisfactorily on OOD test images. The deciding factor here is the diversity of the images from that trial site. A generalization problem may surface if the images from a certain trial site lack sufficient diversity.

In the *scenario 3*, a situation is simulated where the training images are embedded with significant digits representing their true labels. This approach can create a bias in favor of images with overlaid labels. While this bias might produce satisfactory results when tested on images with overlaid labels, it might also prompt the deep network to overlook the broader context of the images, leading to a bias towards the overlaid labels.

In general, while all shortcuts can be classified as a form of bias, not all biases necessarily lead to shortcuts. More specifically, among all types of biases, those that result in high in-distribution performance and low OOD performance are usually considered shortcuts.

For all experiments conducted in this section, KimiaNet [180] was utilized, adhering to the same hyperparameters as in previous sections, with hospital 3 serving as the hold-out set. This provided a consistent framework for analyzing and interpreting the results.

Exploring *Scenario 3*:

In *Scenario 3*, we experiment with an artificial shortcut by overlaying the true labels directly onto the training images. Training a deep network using this method may result in the algorithm exploiting this obvious "clue" during its learning process, likely leading to the development of decision rules based on this prominent shortcut. This type of shortcut is referred to as a *label bias* in the existing literature [80]. Rather than focusing on the broader and more nuanced features of the tissue context, the network primarily recognizes the overlaid label digit. Consequently, when the network encounters an image without an overlaid label after training, it struggles to make accurate predictions since it hasn't learned to recognize the more substantive decision rules. Instead, it will likely default to the class category that was previously overlaid.

Grad-CAM [196] was used to offer some degree of explainability in these cases, generating heatmaps that highlight the salient areas relevant to the classification. As depicted in Figs. 4.4- 4.5, the Grad-CAM heatmaps illustrate how *scenario 3* causes the deep network to fixate almost exclusively on the overlaid label, neglecting the tissue morphology that should ideally inform the decision-making process.

In essence, the overlay of an image's label (an extreme example of a shortcut) takes precedence over all other content within the image in these instances. The deep network essentially operates as a digit recognizer, making decisions based solely on the digit overlaid on the image. When the overlaid class label is either missing or misleading, the deep network struggles to provide accurate results.

Further tests were carried out using KimiaNet, trained under *scenario 3*, with images that did not have class labels overlaid. The result was that the deep network randomly assigned class labels, resulting in an accuracy of approximately 50%, similar to the randomness of flipping a coin.

Exploring *Scenario 1 and 2*:

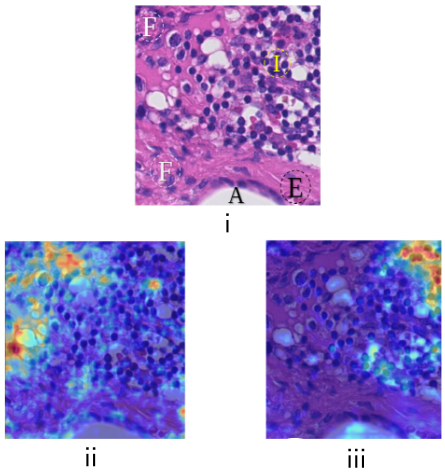


Figure 4.7: **(i)** an OOD healthy patch with different anatomical structures, Ⓘ: Immune cells, Ⓐ: Adipocyte, Ⓕ: Fibrous tissue, Ⓔ: Erythrocyte. **(ii)** GradCAM heatmap for the model trained using *scenario 1* which misclassified the patch as a healthy patch. **(iii)** GradCAM heatmap for the model trained using *scenario 2* which correctly classified the patch.

KimiaNet was trained by segregating images from hospital 3 and utilizing both *Scenario 1 and 2*. Fig.4.6i-ii presents an OOD tumorous patch from hospital 3, which includes pixel-level annotations by a pathologist indicating the tumorous area. The model, when trained using *scenario 2*, correctly identified the image as tumorous, as shown in Fig. 4.6iii which illustrates the salient tumorous area. On the other hand, the model trained under *scenario 1* misclassified the patch as healthy, as evident from the heatmap of salient healthy areas depicted in Fig. 4.6 iv. Although some tumor areas were not highlighted in the explainability heatmap (Fig. 4.6.iii), the activated areas aligned well with the expert annotation, and areas containing lymphocytes were not activated. Conversely, the model trained under *scenario 1* incorrectly classified the patch as healthy, as shown in Fig. 4.6iv which depicts its heatmap for salient healthy regions. This heatmap aligns strongly with the healthy area as per Fig. 4.6.ii, yet certain tumorous regions were incorrectly activated. These regions could be linked to shortcut opportunities that were eliminated by applying the transformations in *scenario 2*.

Fig. 4.7i shows a patch consisting of healthy tissue. The network trained under *scenario 1* wrongly classified this image as tumorous, while the model trained under *scenario 2* correctly identified it as healthy. Fig. 4.7ii and Fig. 4.7iii show heatmaps for salient tumorous and healthy areas for *scenario 1* and *scenario 2*, respectively. It can be observed that the model trained under *scenario 1*, which can be considered as a shortcut-trained model, has incorrectly associated fibrous tissues with the tumorous region. However, in the model trained under *scenario 2*, the prominent healthy areas are mostly composed of immune cells and adipocytes.

From this, it becomes clear that training using *scenario 2* guides the deep network to focus less on nonsemantic features, such as those induced by inconsistent staining colors or variances in morphology and tumor staging across different hospitals or trial sites, and more on what is intended, discerning the semantics of tumorous or healthy patterns.

4.3.2 Variations in Pre-Training: Distinct Aspects of Images

– **Pre-training on a related task vs. ImageNet** – Although pre-training on natural images, including vanilla, Semi-SL, and Semi-WSL ImageNet pre-trained weights, has been the prevalent approach for numerous computer vision tasks, there is substantial evidence suggesting that domain-specific pre-trained weights might prove more effective for certain specialized tasks [119, 225]. This implies that a model pre-trained on an extensive histopathology task, for instance, cancer subtyping on TCGA, could potentially outperform a model with ImageNet pre-training on a histopathology downstream task such as differentiating between tumorous and non-tumorous breast tissues in the CAMELYON dataset.

Histopathological images are distinct in that they exhibit unique variations in cell structures and tissue patterns, attributes which may not be adequately represented in ImageNet, a dataset primarily composed of natural images. By pre-training on the TCGA dataset, the model is primed to recognize and learn features and patterns more relevant to histopathology tasks, thereby improving performance.

In addition, KimiaNet has been trained on a wide variety of common cancer types in various hospitals, including MSKCC, NCI, among others. Employing ERM with the labels representing cancer subtypes, the resultant trained representations can be viewed as somewhat hospital-invariant. The variability among hospitals can function as a form of data augmentation, indirectly enhancing the generalizability of KimiaNet. Consequently, pre-training on the TCGA dataset could result in the model being more proficient at "overlooking" certain irrelevant, hospital-specific aspects of the images, thereby performing better than models pre-trained on ImageNet.

To substantiate this hypothesis, I leveraged heatmaps generated using XAI techniques, which provided a deeper understanding and visual insight into how pre-training on different datasets affected the model's focus and performance.

– **Analyzing Heatmaps of Initial Layers** – Heatmaps generated using XAI techniques, particularly Grad-CAM in this study, tend to emphasize lower-level features such as edges and corners when focused on the initial layers. These contrast with deeper layers, which encapsulate more abstract high-level features. When fine-tuning a suitably pre-trained model for a specific task, these initial layers are typically kept unchanged. This is because they have already been trained to detect "useful" features that are likely to be relevant to the new downstream task [98]. Therefore, a crucial factor in evaluating the suitability of pre-trained weights for downstream tasks is to assess if fine-tuning significantly alters the initial layers.

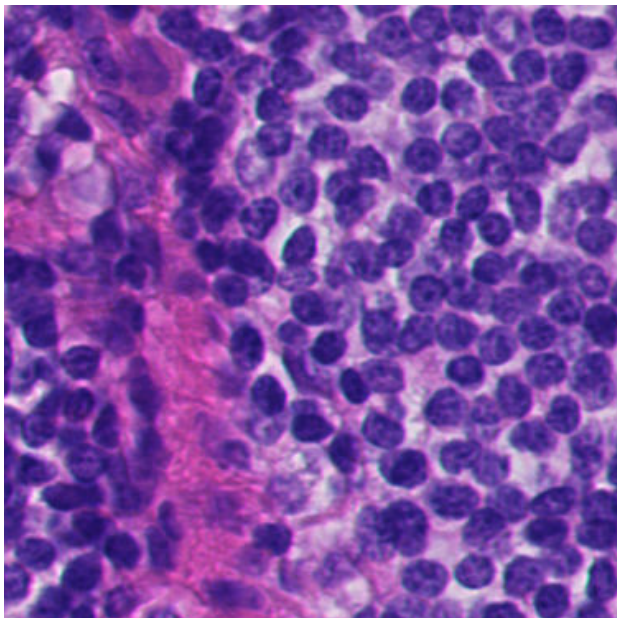


Figure 4.8: Sample non-tumorous patch at 20× magnification from Hospital 3.

In this study, I delve into this issue by inspecting Grad-CAM heatmaps of the image depicted in Fig. 4.8, generated for the first layer of each pre-trained model before and after fine-tuning (see Fig. 4.9). The findings indicate that *the responses of KimiaNet's initial layer remain stable after fine-tuning* when exposed to an OOD healthy patch from hospital 3. This implies that the features captured by this pre-trained model are well-adapted to the downstream task.

In contrast, for other pre-trained models, noticeable changes were observed in the responses of the initial layers, with the model starting with random weights showcasing the most dramatic transformations. These results underline the need for judicious selection of pre-trained weights when preparing models for specific downstream tasks, as the initial assumptions significantly impact the model's learning trajectory and overall performance.

4.4 Final Thoughts

While a predetermined diversification of images, akin to *scenario 2* in this study, may boost OOD generalization in some instances, this is not an absolute rule. I demonstrated that, in certain scenarios, data diversification might paradoxically lead to poor OOD and in-distribution performance by adding complexity to the training of deep learning networks. Consequently, it is not always feasible to anticipate a policy that suits every situation unless the target test data and its distribution are known or accessible. An interesting example of this is learnable augmentation policies [89] employing Cycle-Generative Adversarial Networks (Cycle-GAN)s [267]. These are used to adapt the target data to source data, thereby

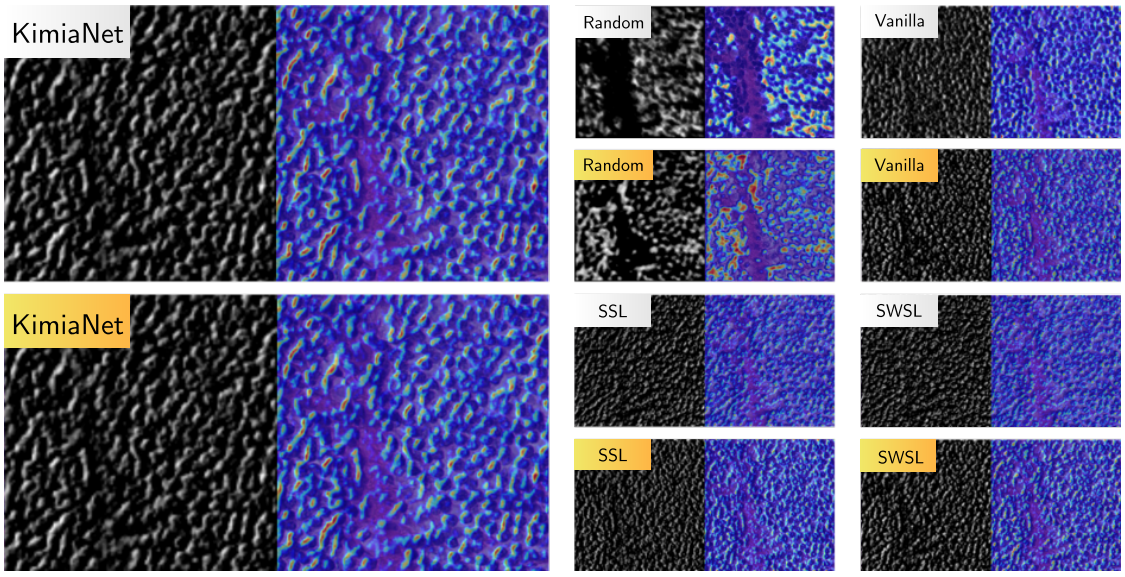


Figure 4.9: Activation maps of first layer weights: pre-trained weights (Gray-highlighted) and fine-tuning (Yellow-highlighted) using the same downstream task for each pre-training scenario.

improving OOD generalization. However, in this study, I made the assumption that target data is not accessible during the training phase.

While some studies have questioned the effectiveness of pre-training, my research underscores the value of employing pre-training in computer vision tasks. I demonstrated that the recent advancements in pre-trained vision models (such as Semi-WSL and Semi-SL) can enhance performance across various scenarios, corroborating findings from other research [261]. More notably, I highlighted that KimiaNet, a pre-trained model specifically tailored to histopathology, can significantly outperform models pre-trained on natural images, particularly when dealing with substantial distribution shifts in the field of histopathology.

Throughout this study, I utilized XAI techniques to elucidate the findings and offer interpretations for certain conclusions. I provided empirical evidence showing that data diversification can boost OOD performance by eliminating shortcuts. Additionally, I explored how the appropriateness of different pre-trained models influences the activation maps of initial layers in deep networks. While some of the conclusions might seem self-evident, this research offers an exhaustive examination of various histopathology trial site repositories, pre-trained models, and image transformations. Furthermore, this study can serve as a valuable reference for practitioners unfamiliar with current concepts and trends in the field. I observed a common practice among the computational pathology community to utilize ImageNet pre-trained models for their histopathology downstream tasks. These findings invite a reconsideration of this practice, highlighting the potential advantages of

using domain-specific pre-trained models.

– Recognizing Limitations –

Despite the extensive examination of various pre-trained models' performance on OOD test data within histopathology repositories in this study, it is paramount to recognize its inherent limitations. Primarily, the scope of the study was confined to the application of ERM on a variety of pre-trained models, while other potential methodologies such as domain adaptation and domain generalization, which could exhibit improved generalization on OOD data, were not explored. This limitation points to potential future research directions aimed at leveraging these alternative approaches for enhancing model generalization.

Secondarily, while the use of XAI techniques was incorporated into the study to interpret the results, the resulting explanations were not subjected to an exhaustive analysis. A more thorough exploration of these explanations has the potential to provide deep insights and contribute to a deeper understanding of the underlying factors that influence the observed distribution shifts in the histopathology domains.

Moreover, the study was limited in its selection of pre-trained models, with attention given exclusively to the vanilla ImageNet, Semi-SL, Semi-WSL, and KimiaNet models. Given the vast array of pre-trained models specifically designed for a multitude of disparate tasks, it is important to note that the results of the study may not be generalized to all pre-trained models. This limitation underscores the need for further research in order to extend and substantiate the findings across a broader range of pre-trained models within the field.

4.5 Summary

In this chapter, a thorough investigation was conducted to examine the performance of various convolutional pre-trained models on OOD test data. These datasets were specifically from unobserved domains during the training phase on histopathology repositories attributed to different trial sites. A myriad of factors, such as trial site repositories, pre-trained models, and image transformations, were meticulously analyzed.

The findings of this research also included a comparative study between models trained entirely from scratch (i.e., without pre-training) and those that had been pre-trained. Specifically, the OOD performance of pre-trained models on natural images was evaluated. This encompassed (1) vanilla pre-trained ImageNet, (2) Semi-SL, and (3) Semi-WSL models pre-trained on IG-1B-Targeted. Furthermore, the performance of a histopathology model (KimiaNet) trained on the most comprehensive histopathology dataset TCGA was studied.

It was discovered that, while the Semi-SL and Semi-WSL pre-trained models provided better OOD performance compared to the vanilla ImageNet pre-trained model, the histopathology pre-trained model exhibited superior overall performance. From the perspective of top-1 accuracy, it was found that diversifying the images in the training set

using suitable image transformations effectively prevented the learning of shortcuts when the distribution shift was significant.

Additionally, the use of XAI techniques enabled high-quality, human-understandable explanations of AI decisions. These XAI techniques were leveraged to perform further investigations, aiding in a deeper understanding of the model's performance. This chapter thus provided significant insights into the OOD performance of various pre-trained models on histopathology datasets and shed light on effective strategies to improve their generalization capability. These findings lay the groundwork for the subsequent chapters of this thesis, where further techniques to enhance model generalization will be explored.

In the following chapter, the idea of employing techniques from DG is proposed as a more reliable alternative to ERM. This approach aims to address and ameliorate the existing shifts and gaps between various hospital repositories.

Chapter 5

Hospital-Agnostic Image Representation in Digital Pathology

Prologue

The content of this chapter is based on an article published during my Ph.D. research: Hospital-Agnostic Image Representation Learning in Digital Pathology- M Sikaroudi, S Rahnamayan, HR Tizhoosh- IEEE Engineering in Medicine & Biology Conference (EMBC) 2022 [208]

5.1 Motivation

This chapter presents a novel methodology for improving the generalization of computational pathology models by extracting domain or hospital-invariant features. Computational pathology has been revolutionized by the digitization of biopsy slides and the application of ML for image classification and cancer diagnosis. Despite these advancements, the utility of these models in clinical practice remains questionable due to the scarcity of expert-labeled training data and their limited ability to generalize beyond this data.

Transfer learning is a potential solution to these limitations, given its ability to repurpose previously learned abstract information in new contexts. However, deep learning networks often fail to account for OOD scenarios and differences between the source and target domain distributions, a problem known as "domain shift". This shift is common in histopathology due to variations in slide preparation, staining procedure, and scanner characteristics across different trial sites.

In this context, training robust models that are invariant to domain shifts (also known as "hospital-agnostic") is crucial for accurate diagnoses and ultimately, patient well-being. While increasing training data diversity or normalizing stains could potentially alleviate this problem, these methods are often subjective and expensive.

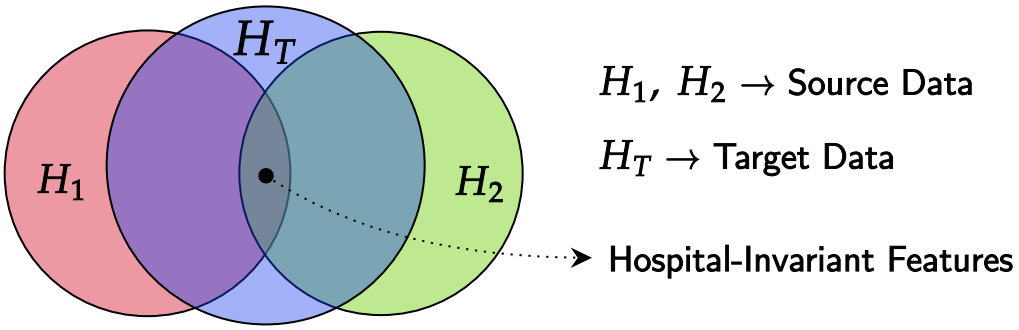


Figure 5.1: The Venn diagram illustrates the feature space of H_1 and H_2, which represent disparate source data, as well as H_T, denoting the target data. Features that remain consistent across all hospitals are also depicted.

To address these challenges, the chapter proposes the use of MDL regimes and a specific DG technique called MASF. The aim is to take advantage of these methods to extract domain-invariant features Fig.5.1, thereby enhancing the models' ability to generalize across different hospitals and conditions. This innovative approach offers a promising avenue to improve the practical utility of computational pathology models.

5.2 Methodology

5.2.1 Preprocessing

Suppose we have K collections of WSI repositories, each of which has been gathered from a distinct trial site. We denote these as $\mathcal{H}_s = \{H_k\}_{k=1}^{K}$, where each H_k represents the WSI repository from the k^{th} trial site.

Our first step involves the segmentation of tissue regions from the background regions in each of the WSIs. For this purpose, we employ the Otsu algorithm [165], a popular method for image thresholding. Once the segmentation is complete, we have distinct tissue regions that can be better analyzed.

Following this segmentation process, we proceed to generate a dataset of patches. Each of these patches is a piece of the segmented tissue region, and this process results in a transformed version of our original set of WSI repositories, which we now denote as $\mathcal{H}_p = \{H_k\}_{k=1}^{K}$. These patches are what will be fed into the CNN in the subsequent steps of our methodology.

In order to evaluate the model, we employ the leave-one-*hospital*-out technique, as detailed in Algorithm 2. This technique is a variation of the widely adopted practice

of leave-one-out cross-validation but adapted to our context, with the "hospital" or trial site acting as the unit of analysis. Under this approach, we train our model on the WSI repositories collected from $K-1$ trial sites while holding out a single repository for testing.

The advantage of this method lies in its realistic evaluation of the model's performance. By testing on unseen data from a different trial site, we gain valuable insights into how well our model can generalize and apply its learned knowledge to new, distinct data. This can be particularly beneficial in a field such as computational pathology, where models need to be capable of accurately analyzing data from varied sources.

Following the extraction and preparation of the patches from the WSI repositories, we denote this comprehensive dataset as $\mathcal{H}_p$. To enable a robust testing and training protocol for our model, this dataset is then partitioned into two distinct subsets, namely $\mathcal{H}_{\text{external}}$ and $\mathcal{H}_{\text{internal}}$.

The subset $\mathcal{H}_{\text{external}} = \{H_k\}$ signifies the hold-out repository, which is set aside explicitly for the purpose of testing. This subset plays a critical role in the invented approach as it represents 'unseen' data for the model. The intention behind this strategy is to evaluate the performance of our model on fresh, previously unseen data, a scenario that closely replicates how the model will operate in real-world conditions. This provides a realistic and unbiased assessment of the model's ability to generalize and adapt to new information.

On the other hand, the subset $\mathcal{H}_{\text{internal}} = \{H_i\}_{i \neq k}^{K}$ encompasses the remaining repositories, and these are utilized to form the training dataset for the model. This comprehensive set of patches drawn from multiple trial sites serves as fertile ground for the model to learn and recognize a variety of complex tissue patterns. By learning from this diverse set, the model can gain a broader and more comprehensive understanding, thereby enhancing its predictive capabilities and ability to generalize across different contexts.

Our method of splitting the dataset into separate training and testing subsets allows us to perform a comprehensive evaluation of our model's performance. By testing the model on fresh, unseen data, we ensure an unbiased assessment that closely mirrors real-world scenarios. The ultimate goal of this methodology is to create a model that is not only accurate but also possesses a reliable capacity for processing and interpreting histopathological images from various hospitals. This could significantly increase its usability and value in the field of computational pathology.

5.2.2 Hospital-Agnostic Learning Regime

Within the framework of our learning regime, let's consider that there are K source domain trial sites designated for training, represented by $\mathcal{H}_{\text{internal}} = \{H_i\}_{i \neq k}^{K}$. It's worth noting that these trial sites consist of different hospitals, each with its own distinct set of WSI repositories.

In every iteration of the learning process, we further divide these source domain trial sites into two subsets - one for meta-training and the other for meta-testing, which we

denote as $\mathcal{H}_{\text{tr}}$ and $\mathcal{H}_{\text{te}}$, respectively. This approach is similar to the typical train-test split in machine learning, but we incorporate a hierarchical level, commonly referred to as 'meta', to enhance the generalization of the model.

For the implementation of the MASF technique, we utilize three different loss functions. These loss functions are critical components of the learning process, guiding the model in its task of learning from the data. The following sections will elaborate on these loss functions and their role in the learning regime.

Cross-Entropy Loss

$$\mathcal{L}_{\text{ce}}(\mathcal{H}_{\text{tr}}; \psi, \theta) := \frac{-1}{|\mathcal{H}_{\text{tr}}|} \sum_{\mathcal{H} \in \mathcal{H}_{\text{tr}}} \frac{1}{|\mathcal{H}|} \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{H}} \sum_{c=1}^{C} \mathbb{I}(y = c) \log \mathbb{I}(\widehat{y} = c), \tag{5.1}$$

$|\cdot|$ indicates the cardinality of the set, C is the number of classes, y is the true label for $\mathbf{x}$, $\widehat{y}$ is the predicted label for $\mathbf{x}$, and $\mathbb{I}(\cdot)$ is the indicator function which is one when its condition is met, and zero otherwise.

Hospital-Alignment Loss

For two hospitals $\mathcal{H}_i$ and $\mathcal{H}_j$ datasets, the Hospital-alignment loss averaged over all the C classes, is calculated as,

$$\ell_{\text{hospital alignment}}(\mathcal{H}_i, \mathcal{H}_j; \psi', \theta') := \frac{1}{C} \sum_{c=1}^{C} \frac{1}{2} \Big[\mathrm{D}_{\mathrm{KL}}(\mathbf{s_c^{(i)}} \| \mathbf{s_c^{(j)}}) + \mathrm{D}_{\mathrm{KL}}(\mathbf{s_c^{(j)}} \| \mathbf{s_c^{(i)}}) \Big], \tag{5.2}$$

where D_{KL} denotes the symmetrized Kullback–Leibler divergence, and $\mathbf{s_c^{(\cdot)}}$ denotes the soft confusion matrix which is calculated by applying softmax function with temperature $\tau > 1$ on the output of classification subnetwork, i.e. S_θ.

Triplet Loss

Triplet loss [193, 204] is used for promoting hospital-independent class-specific cohesion and separation of instances. Each anchor, positive, and negative instances is denoted by

$\mathbf{x_a}$, $\mathbf{x_p}$, and $\mathbf{x_n}$, respectively. For a batch of triplets, $\tau := \{\mathbf{x_a^b}, \mathbf{x_p^b}, \mathbf{x_n^b}\}_{\mathbf{b=1}}^{\mathbf{B}}$, from all the source domain datasets $\{\mathcal{H}_k\}_{k=1}^{K}$, the average triplet loss is

$$\begin{aligned}\mathcal{L}_{\text{triplet}}(\tau; \psi', \phi) := \\ \frac{1}{B}\sum_{b=1}^{B}\Big[\|M_\phi(G_{\psi'}(\mathbf{x_a^b})) - \mathbf{M}_\phi(\mathbf{G}_{\psi'}(\mathbf{x_p^b}))\|_{\mathbf{2}}^{\mathbf{2}} \; - \\ \|M_\phi(G_{\psi'}(\mathbf{x_a^b})) - \mathbf{M}_\phi(\mathbf{G}_{\psi'}(\mathbf{x_n^b}))\|_{\mathbf{2}}^{\mathbf{2}} + \alpha\Big]_+,\end{aligned} \tag{5.3}$$

where $\|\cdot\|_2$ is the ℓ_2 norm, α is a margin and $[\cdot]_+ := \max(\cdot, 0)$.

5.2.3 Gradient Updating

First ψ and θ weights of S_θ and G_ψ are updated given by:

$$(\psi', \theta') \leftarrow (\psi, \theta) - \alpha\nabla_{\psi,\theta}\mathcal{L}_{\text{ce}}(\mathcal{H}_{\text{tr}}; \psi, \theta), \tag{5.4}$$

where α is the learning rate and $\nabla_{\psi,\theta}$ indicates the gradient with respect to ψ and θ parameters. Then, the meta loss is calculated using weighted sum of hospital-alignment and triplet losses as

$$\begin{aligned}\mathcal{L}_{\text{meta}}(\mathcal{H}_{\text{tr}}, \mathcal{H}_{\text{te}}, \tau; \psi', \theta', \phi) \leftarrow \\ \beta_1\mathcal{L}_{\text{hospital alignment}}(\mathcal{H}_{\text{tr}}, \mathcal{H}_{\text{te}}; \psi', \theta') \quad + \\ \beta_2\mathcal{L}_{\text{triplet}}(\tau; \psi', \phi),\end{aligned} \tag{5.5}$$

where β_1 and β_2 are positive. After Eq. 5.4, two other gradient descent steps are done as

$$(\psi, \theta) \leftarrow (\psi, \theta) - \eta\nabla_{\psi,\theta}(\mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{meta}}), \tag{5.6}$$

$$\phi \leftarrow \phi - \gamma\nabla_\phi\mathcal{L}_{\text{triplet}}(\tau; \psi', \phi), \tag{5.7}$$

where η and γ are the learning rates and ϕ is the M_ϕ subnetwork parameters. Eqs. (5.4, 5.6, and 5.7) are repeated iteratively until convergence.

5.3 Experiments

5.3.1 Dataset

Renal Cell Carcinoma — RCC is the most common form of kidney cancer observed in adults. It presents itself as a diverse group of diseases, each distinguished by their unique morphology, molecular traits, clinical consequences, and responses to treatment. Since

Algorithm 2: Hospital-Agnostic Approach

Data: There are K sets of WSIs: $\mathcal{H}_s = \{H_k\}_{k=1}^{K}$, and hyperparameters $\alpha, \eta, \gamma, \beta_1, \beta_2$

Result: There will be K sets of feature extractor G_ψ, classifier S_θ, metric embedding M_ϕ subnetworks

Preprocessing:

1. Segmentation of tissues from the background using Otsu method [165]
2. Extracting 227×227 patches from the foreground regions
3. Create $\mathcal{H}_p = \{H_k\}_{k=1}^{K}$ using the patches

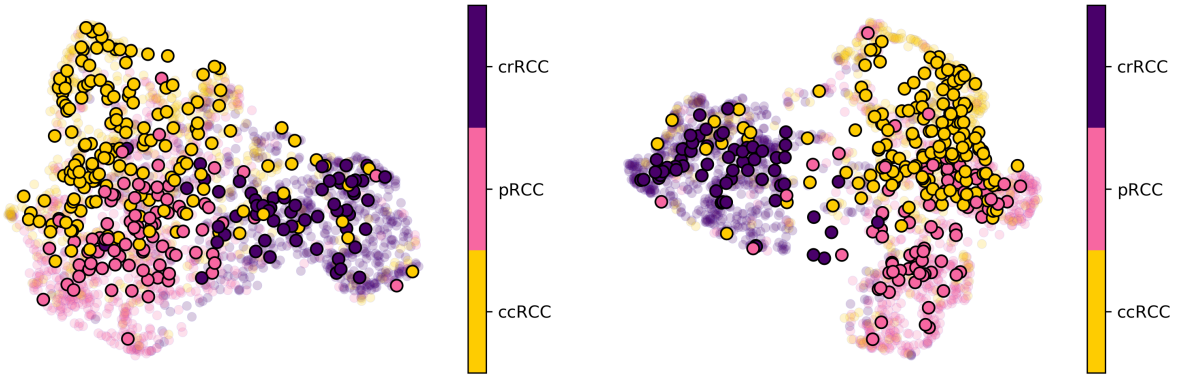
foreach *k, splits* $\mathcal{H}_p$ *into* $\mathcal{H}_{external} = \{H_k\}$ *and* $\mathcal{H}_{internal} = \{H_i\}_{i \neq k}^{K}$ **do**
  repeat
    Randomly split $\mathcal{H}_{\text{internal}}$ into disjoint meta-train $\mathcal{H}_{tr}$ and meta-test $\mathcal{H}_{te}$,
    Update using Equation 5.4,
    Compute hospital alignment loss:
    $\mathcal{L}_{\text{hospital alignment}} \leftarrow \frac{1}{|\mathcal{H}_{\text{tr}}|} \sum_{\frac{1}{|\mathcal{H}_{\text{te}}|} \mathcal{H}_i \in \mathcal{H}_{\text{tr}}} \sum_{\mathcal{H}_j \in \mathcal{H}_{\text{te}}} \ell_{\text{hospital alignment}}(\mathcal{H}_i, \mathcal{H}_j; \psi', \theta')$,
    Compute triplet loss using Equation 5.3,
    Compute meta loss using Equation 5.5,
    Update using Equation 5.6,
    Update using Equation 5.7,
  until *convergence*;
end

RCCs are classified on the basis of their histological subtypes, the task of classification is central to diagnosis and is key to increasing the probability of successful treatment. Given its importance and the complexity involved in its classification, we've chosen RCC as our case study.

WSI repository — The WSIs and relevant clinical information for this study was sourced from TCGA data portal. Among the complete set of data, some WSIs were excluded due to problems with readability and compatibility. Moreover, for the purpose of this research, I focused solely on diagnostic slides that were scanned at a 40x magnification.

In selecting trial sites for the study, our criteria necessitated a substantial number ($\gtrapprox 30$) of WSIs spanning all three RCC subtypes (Clear Cell Renal Cell Carcinoma (ccRCC), Papillary Renal Cell Carcinoma (pRCC), Chromophobe Renal Cell Carcinoma (crRCC)). From the TCGA, only three sites fit this requirement: the NCI, the IGC, and the MSKCC. To add a fourth center, we combined the repositories from Harvard and MD Anderson cancer centers (HMD).

Taken together, our study utilized a total of 467 RCC WSIs from the TCGA, which provided us with a rich and diverse set of data to train and test our models. This compre-



(a) (b)

Figure 5.2: (a) ERM, (b) Hospital-agnostic. The hold-out trial site is "**NCI**". Note that the resulting 2-dimensional representations have been transparently visualized for each patch representation by its ground-truth slide-level label. The 2-dimensional representations of all patches were aggregated (averaged) for each WSI to attain the slide-level representations which are shaded opaque with a dark border.

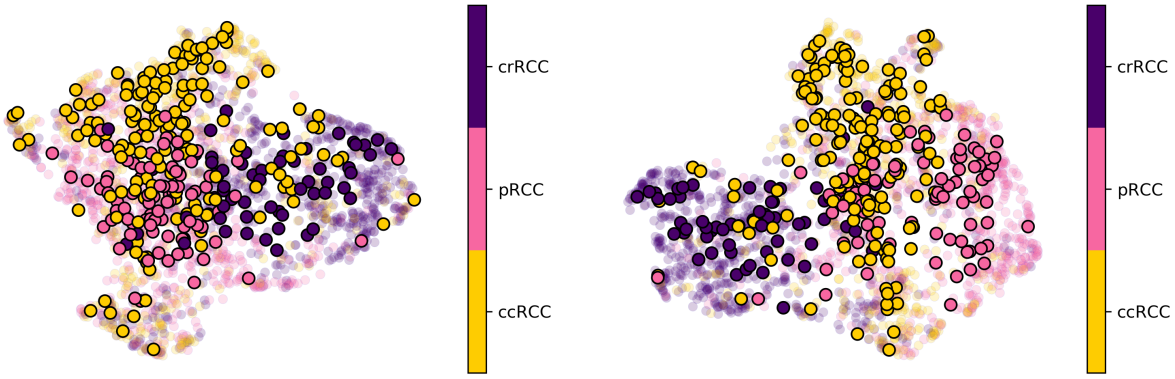
hensive dataset will help us to develop a robust and effective model for classifying RCC, thereby contributing to advancements in the field of computational pathology.

Details — Each WSI repository was divided into chunks of 45%, 45%, and 10% (akin to [132]), allocated respectively for training, validation, and testing. Following this division, the foreground regions of each WSI, representing the tissue, were extracted. Subsequently, a process of morphological closing was applied to each WSI to fill in any minor gaps and holes.

In order to ensure compatibility with the input size requirements of the backbone network (AlexNet), $227 \times 227 \times 3$ RGB patches were extracted from the segmented tissue of each WSI at a 40x magnification without any overlap. Any patches that contained more than 50% of the background region were discarded. Furthermore, a re-sampling of all the extracted patches was conducted to create a balanced dataset. For each of the trial sites and subtypes in $\mathcal{H}_p$, approximately 70,000 patches were sampled.

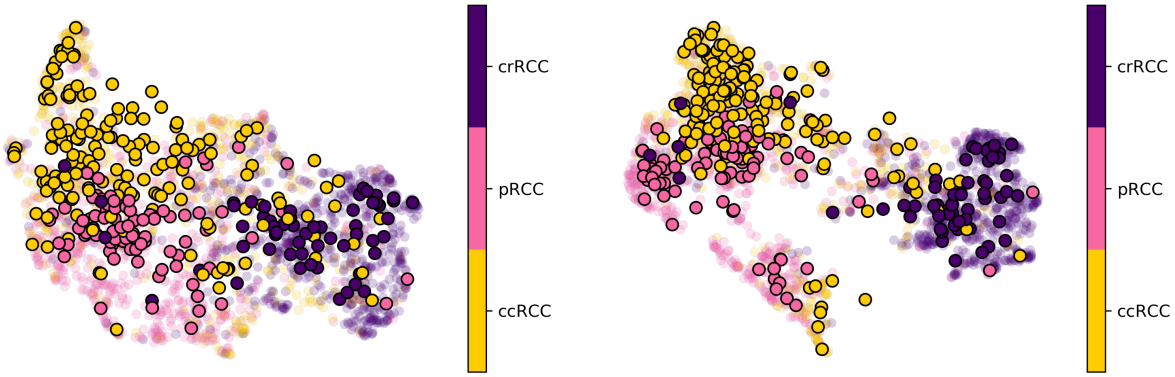
5.3.2 Experimental Setup

Baseline – In the context of these experiments, the baseline method was established by fine-tuning the backbone architecture using traditional cross-entropy loss, also known as ERM. This baseline was constructed by amalgamating all the WSIs sourced from various trial sites, followed by training the model $G_\psi \circ S_\theta$ using standard supervised learning based on $\mathcal{L}_{\text{ce}}$. It's important to note that the same hyperparameters were maintained as those used in the hospital-agnostic regime, ensuring consistency across different aspects of the experiment.



(a) (b)

Figure 5.3: (a) ERM, (b) HA. The hold-out trial site is "**MSKCC**".



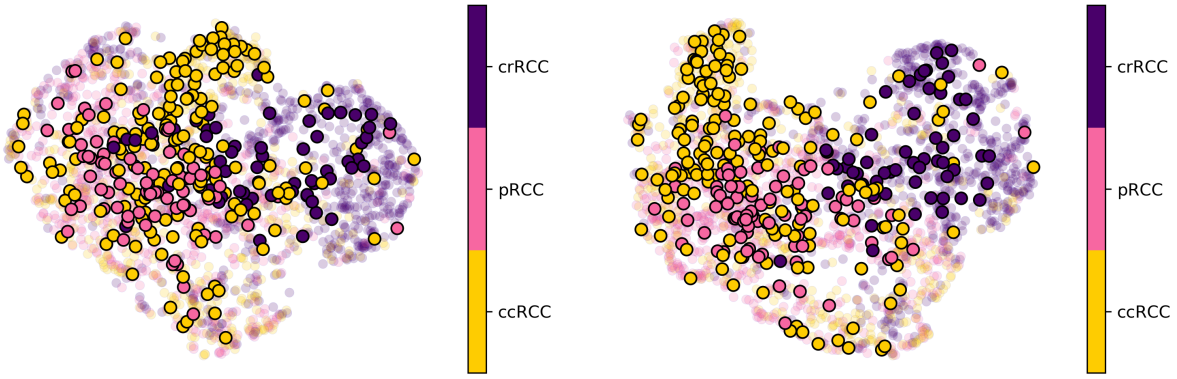
(a) (b)

Figure 5.4: (a) ERM, (b) HA. The hold-out trial site is "**IGC**".

Backbone— The proposed method in this study is designed to be model-agnostic and consequently, architecture-agnostic. This essentially means that general architecture remains a universal concept, making it possible for the core idea to be implemented across a wide range of architectures. In the case of this study, the backbone of choice was "AlexNet", which had been pre-trained on the ImageNet dataset.

Hyperparameters and Details— The Adam optimizer [115], initialized with a learning rate of 10^{-3}, was utilized for the optimization process. Two fully-connected layers, with output sizes of 1,024 and 256, were stacked together to form the subnetwork for metric loss, M_ϕ, which was then connected to the final fully-connected layer. For the calculation of $\mathcal{L}_{\text{triplet}}$, the triplet loss was employed with $\beta_2 = 0.005$ and $\beta_1 = 1.0$, in order to achieve a scale similar to that of $\mathcal{L}_{\text{ce}}$ and $\mathcal{L}_{\text{hospital alignment}}$. To prevent gradient explosion during the inner optimization process, gradients with a norm exceeding a predefined threshold[1]

[1] As per [52], the threshold was set at 2.0



(a) (b)

Figure 5.5: (a) ERM, (b) HA. The hold-out trial site corresponds to HMD.

were clipped. This step involved the use of a simple, non-adaptive gradient descent approach with a learning rate $\alpha = 1e - 5$. Additionally, an Adam optimizer was used for meta-updates with a learning rate of $\eta = 10^{-5}$, no decay, and a batch size of 432 patches. The metric-learning margin's hyperparameter was set heuristically at 10 (following the guidelines in [52]), which was based on the evaluation of distances between clusters of class characteristics. Lastly, the learning rate for metric loss was assigned a value of $\gamma = 10^{-5}$ with a maximum number of iterations capped at $1,000$.

5.3.3 Results

In the context of this study, a leave-one-*hospital*-out approach was implemented for each trial site, which is to say that $\mathcal{H}_{\text{external}}$. This meant that the backbone was trained exclusively using the remaining repositories, denoted as $\mathcal{H}_{\text{internal}}$.

Low-dimensional Embedding Visualization— Upon completion of the training, the model was utilized to derive a 4,096-dimensional latent space representation for every patch present in the corresponding hold-out test set. To achieve a lower-dimensional representation that is easier to visualize, we followed the best practice as suggested in the existing literature [189]. Specifically, the first 20 principal components were obtained using Principal Component Analysis (PCA) [251], after which Uniform Manifold Approximation and Projection (UMAP) [153] was applied to reduce the dimensionality to 2. As can be observed in Figs. 5.2, 5.3, 5.4, and 5.5, the hospital-agnostic learning approach demonstrated superior performance over the baseline (ERM) in terms of producing a more discriminative embedding space.

RCC Classification Accuracy— Once the training process was completed, slide-level accuracy was computed. This involved averaging the softmax outputs - probability scores that indicate the degree of belongingness to each of the RCC subtypes (ccRCC, pRCC, and crRCC) - over the patches of each WSI. Consequently, a predicted subtype could

Table 5.1: Slide-level accuracy for different trial sites.

	Accuracy(%)	
Hold-out Trial Site	HA	ERM
IGC	79.31	**80.45**
HMD	**79.31**	72.41
MSKCC	**82.65**	81.18
NCI	**84.09**	81.81

be assigned to each WSI. The results, as outlined in Table 5.1, show that the proposed hospital-agnostic learning regime outperformed the baseline when the hold-out trial sites were "NCI", "Harvard and MD Anderson", and "MSKCC". A notable example can be seen when the "Harvard and MD Anderson" was the trial site. Since this repository was created by combining two distinct trial sites, a more significant domain shift was expected. As per the results in Table. 5.1, the hospital-agnostic learning regime was successful in surpassing the baseline by approximately 7% when the hold-out repository was "Harvard and MD Anderson".

5.4 Conclusion and Summary

In this chapter, I proposed an innovative hospital-agnostic learning regime aimed at enhancing the generalizability of computational models in medical applications, such as in diagnosing RCC. This approach was inspired by the concept of DG, with a specific focus on MASF, a DG technique. The underlying concept here involves training a DNN through an episodic learning regime that utilizes three distinct loss functions.

The first of these is the cross-entropy loss, employed for achieving hard class separation, which essentially helps the model to distinctly categorize different classes in a manner that reduces misclassification. The second loss function is the triplet loss, designed for soft class separation. This loss function aids in embedding relative distances between samples from the same class and different classes, thereby promoting finer, nuanced separation within the learned feature space.

The third and final loss function is the KL divergence, used specifically for hospital alignment. This assists in mitigating the distributional discrepancies that may arise due to the unique characteristics of data from different hospitals, thereby aligning the different hospital domains in the model's learned representation.

What makes this approach unique is its emphasis on learning invariant features across various domains or hospitals. This means that the model strives to focus on the common and universal characteristics across different trial sites while marginalizing the site-specific variances that could potentially hinder its generalization capacity. By promoting the learning of such invariant features, I aspire to develop a model that can accurately and reliably

interpret and process histopathological images, regardless of their source hospital. The effectiveness of this approach in achieving a discriminative latent space representation and classification of RCC subtypes was demonstrated and analyzed. This analysis was performed through low-dimensional embedding visualization and classification accuracy and these were compared with ERM.

While emphasizing domain- or hospital-invariant features paves the way for a more generalized model, as I'll discuss in the next chapter, there exists an additional set of features that can be harnessed to enhance the model's generalization capabilities even further.

Chapter 6

Leveraging All Levels of Feature Abstraction for Improving the Generalization

Prologue

The content of this chapter is based on an article published during my Ph.D. research: ALFA–Leveraging All Levels of Feature Abstraction for Enhancing the Generalization of Histopathology Image Classification Across Unseen Hospitals- M Sikaroudi, M Hosseini, S Rahnamayan, HR Tizhoosh- Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshop, CVAMD, 2023 [209]

6.1 Motivation

In DG, domain alignment techniques are popular [237, 142, 52], as they aim to minimize differences among source domains to learn domain-invariant features that can withstand unforeseen shifts in the target domain [219].

Ignoring domain-specific information in favor of domain-invariant features may not always lead to the best generalization performance, as noted by Mancini et al. [152] and Shankar et al. [200]. Bui et al. [27] proposed the mDSDI method and provided a mathematical proof for that.

Figure 6.1 illustrates that, apart from the invariant features, there exists a set of unique features specific to each hospital. These features are of significant importance as they can establish a mapping from the feature space to the label space.

In addition to invariant and specific features, there exists a distinct set of features that can be attained without the use of either class labels or domain labels. SSL encourages

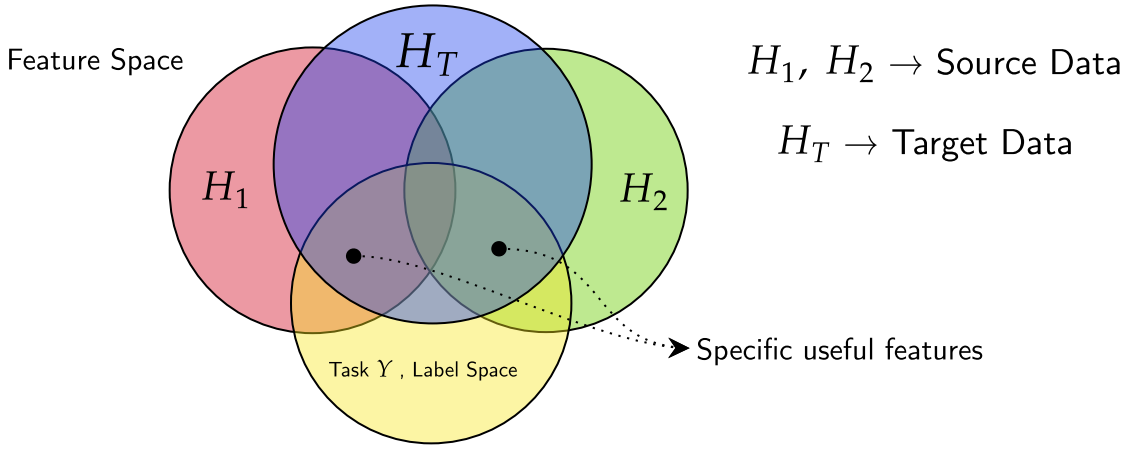


Figure 6.1: The Venn diagram delineates the feature space of the source hospitals (H_1 and H_2) in conjunction with that of the target hospital (H_T). The yellow area demarcates the label space employed in the classification task.

the model to learn features that are relevant for predicting the data itself without being constrained by any dominant hard labels [170, 36]. This can result in model learning features that capture more basic perceptual information about the data, such as edges, corners, and textures, in digital images [31]. By combining self-supervision representations with invariant and specific representations, a range of representations can be obtained that encompass all levels of *feature abstraction*.

This chapter proposes a new method that extracts disentangled feature abstractions at all levels to enhance the model's ability to generalize to new data from different hospitals/domains. Accordingly, the main contributions are:

- My proposed approach, called ALFA, is an extension of the mDSDI technique for DG. ALFA disentangles the components of SSL, domain-invariant, and domain-specific representations to reduce redundancy and improve generalization to unseen target data. By exploiting all levels of feature abstraction, ALFA strives to fully utilize the available information in the dataset.

- The mDSDI [27] approach utilizes adversarial training to extract domain-invariant features, but it can be unstable due to a non-differentiable step (gradient reversal). Therefore, in this chapter, a loss function called "*soft class-domain alignment*" is proposed to minimize the average divergence between two domain probability distributions and a target probability distribution representing a soft class label for each class. This loss function provides better stability during optimization and more distinct latent space for the representations.

- To evaluate the effectiveness of the proposed improvement, I conduct experiments on two public datasets: the PACS [133] benchmark for DG and a RCC subtyping task extracted WSIs of TCGA [100] data portal.

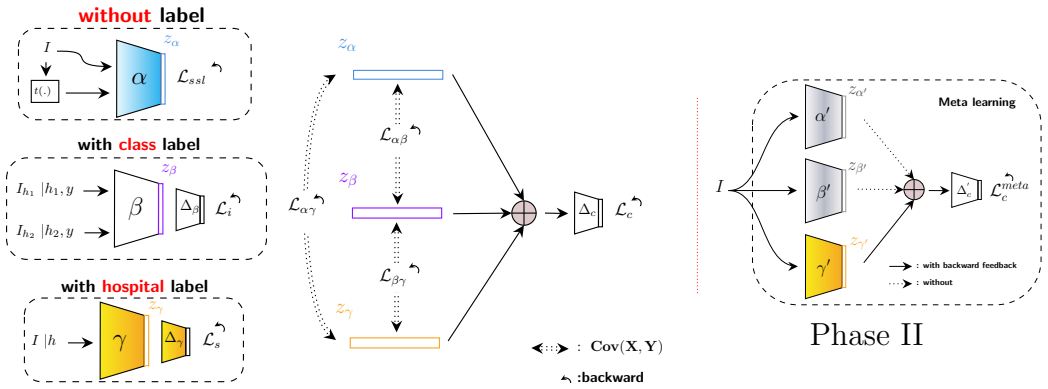


Figure 6.2: ALFA has two phases: In Phase I, three feature extractors extract different levels of feature abstraction, and disentangled features are concatenated for classification. In Phase II, updated feature extractors' representations are concatenated and fed into the updated classifier to update parameters in a Meta-learning fashion while α' and β' feature extractors remain frozen.

6.2 Methods

The key concepts underlying the invented approach have been discussed up to this point. In this section, the specific details of how ALFA is implemented will be delved into. A visual representation and a high-level overview of ALFA framework are provided in Fig. 6.2. It should be noted that the symbols "$\oplus$" and "$\circ$" have been employed as concatenation and union operators, respectively.

Several components are included in the invented integrated network: (1) an SSL representation $z_\alpha^I = \alpha(I;\theta_\alpha))$, with α being the SSL encoder that is parameterized by θ_α; (2) a domain-invariant representation $z_\beta^I = \beta(I;\theta_\beta))$, parameterized by θ_β, which serves as the domain invariant feature extractor; (3) domain-specific representation $z_\gamma^I = \gamma(I;\theta_\gamma))$, where γ stands for the domain-specific feature extractor parameterized by θ_γ; (4) a domain aligner, parameterized by θ_{Δ_β}, denoted as $\Delta_\beta(z_\beta^I;\theta_{\Delta_\beta}) : z_\beta^I \rightarrow \overline{1:N_c}$; (5) a domain classifier, parameterized by θ_{Δ_γ}, represented as $\Delta_\gamma(z_\gamma^I;\theta_{\Delta_\gamma}) : z_\gamma^I \rightarrow \overline{1:N_h}$, where N_c and N_h refer to the number of classes, and the number of participating hospitals/domains in the training, respectively; (6) a regular classifier, parameterized by θ_c, i.e., $\Delta_c(z_\alpha^I \oplus z_\beta^I \oplus z_\gamma^I;\theta_c) : z_\alpha^I \oplus z_\beta^I \oplus z_\gamma^I \rightarrow \overline{1:N_c}$. The hospital and images sample spaces are represented by $\mathcal{H}$ and $\mathcal{I}$ respectively. Images, or I, are denoted by their target labels, or y, and hospital labels, or h, as (I, y, h).

6.2.1 Phase I: Extracting different levels of feature abstraction

Construction of Self-supervision Representations: Capturing the "Most General Features"

At the core of "Phase I" is the training of a SSL feature extractor, represented by $\alpha(I;\theta_\alpha)$. This component is designed to discern and learn generalizable features, which will become instrumental for accomplishing downstream tasks. The learning process employs a technique known as a triplet loss, a methodology similar to the one detailed in the study by Wang et al. [247].

The first step in this learning process involves the creation of pseudo-classes. These are generated by applying a set of transformations, denoted as $\mathcal{T}$, to a single image. This leads to the production of an augmented image, symbolized as $I_t = t(I)|t \sim \mathcal{T}$.

An image originating from a different pseudo-class, denoted as I_d, is also utilized in the training process. Together, these images serve as training inputs for the network, which employs the following triplet loss formula:

$$\mathcal{L}_{\text{SSL}} = \max(||z_\alpha^I - z_\alpha^{I_t}||_2 - ||z_\alpha^I - z_\alpha^{I_d}||_2 + M, 0), \tag{6.1}$$

In this equation, $||.||_2$ signifies the Euclidean distance and M represents the margin, which is set heuristically at 1.5. This formula guides the network's learning process by maximizing the intra-class similarity and minimizing the inter-class similarity. The transformations set $\mathcal{T}$ includes HED jitter with a jitter parameter of $\theta = 0.05$, as prescribed by Tellez et al., 2018 [228]. Additionally, random affine transformations are applied, including rotation (varying between -10 and 10 degrees), translation (ranging from 0 to 0.1), and shear (from -1 to 1) in both x and y directions. These transformations contribute to the generation of diverse pseudo-classes, providing a wide range of inputs to strengthen the learning process. In Eq. 6.1, the margin, represented by M, is set to 1.5. The mining strategy employed is semi-hard mining, for which the margin is specified as 0.7.

Hospital-invariant representations: "general features across hospitals" –

Aligning class relationships across hospitals promotes more transferable knowledge for model generalization compared to individual hard label prediction [52]. This study aims to impose an overall pattern of retrieved features that represent the intrinsic similarity between the semantic structures of different classes. Soft labels and class labels are used for consistency across domains.

– **Soft Confusion Matrix:** Let $\overline{z}_c^{(k)}$ denote the mean of class c in the domain k in the embedding space. I use *softmax* activation for representing the probability of belonging to classes as $s_c^{(k)} = softmax(\Delta_\beta(\overline{z}_c^{(k)})/\tau)$, where $\tau > 1$ is the temperature. The group of

soft labels, $[s_c^{(k)}]_{c=1}^{C}$, serves as a form of *soft confusion matrix* associated with a particular domain/hospital. By combining the class labels with this soft confusion matrix, our both goals, i.e., domain alignment and retaining the class relationship, can be fulfilled.

– **Soft Class Label Injection:** In addition to this soft confusion matrix, proposed in [52], I have injected the soft class labels through a discrete probability density function that represents each class as follows

$$p_c \coloneqq \{i | i = \delta_i^c + (1 - \delta_i^c)(\frac{\zeta}{n_c - 1})\}, \tag{6.2}$$

where δ_i^c is the Kronecker delta which is defined as $\delta_i^c \coloneqq \begin{cases} 1 & \text{if } \ i = c \\ 0 & \text{otherwise} \end{cases}$, and $\zeta \lessapprox 1$ is a constant value indicating a high probability value (e.g., $\zeta = 0.9$), and n_c is the number of classes.

Overall, the aim is to minimize the average divergence [198] over all the C classes between three distributions: two arbitrary hospitals/domains samples drawn from the two training hospitals' images, $(I_{h_1}, ., h_1) \sim \mathcal{H}_1, (I_{h_2}, ., h_2) \sim \mathcal{H}_2$, and the probability distribution of each class, defined in Eq. 6.2.

The *Soft Class-Domain Alignment* loss which serves as the domain-invariant loss in our design is defined as

$$\begin{aligned} \mathcal{L}_{\text{i}}((I_{h_1}, ., h_1), (I_{h_2}, ., h_2); \theta_\beta \circ \theta_{\Delta_\beta}) \coloneqq \frac{1}{C} \sum_{c=1}^{C} \frac{1}{6} \Big[D_{\text{KL}}(s_c^{(h_1)} \| s_c^{(h_2)}) + D_{\text{KL}}(s_c^{(h_2)} \| s_c^{(h_1)}) + \\ D_{\text{KL}}(p_c \| s_c^{(h_2)}) + D_{\text{KL}}(s_c^{(h_2)} \| p_c) + D_{\text{KL}}(s_c^{(h_1)} \| p_c) + D_{\text{KL}}(p_c \| s_c^{(h_1)}) \Big], \end{aligned} \tag{6.3}$$

where $D_{\text{KL}}(p \| q) = \sum_r p_r \log \frac{p_r}{q_r}$, and $\theta_\beta \circ \theta_{\Delta_\beta}$ is the union of all the parameters for β feature extractor and Δ_β classifier.

The inclusion of the SSL features as a level of feature abstraction could raise the question of how distinct SSL features are from the hospital-invariant features. The subsequent theorem and lemma aim to provide clarity on this matter:

Theorem 6.2.1 *Given transformations T_1 for α and T_2 for different hospitals (due to distribution shift), and an optimization objective to minimize the covariance between z_α and z_β obtained respectively by T_1 (explicit data augmentation) and T_2 (implicit changes due to sources of distribution shifts), are distinct and uncorrelated in the feature space and both contribute unique information for mapping from feature space to label space.*

Lemma 6.2.2 *The following assertions hold for the features $z_\alpha = \alpha(T_1(I))$ and $z_\beta = \beta(T_2(I))$ for a given image I:*

1. *z_α and z_β are uncorrelated:*

   If the optimization objective successfully minimizes the covariance, the covariance between z_α and z_β, $Cov(z_\alpha, z_\beta)$, will be close to zero. This indicates that z_α and z_β are uncorrelated, i.e., changes in z_α do not predict changes in z_β and vice versa.

2. *z_α and z_β contribute unique information to the mapping:*

   Let us consider a mapping function M that maps the feature space to the label space. For z_α and z_β, the mapping function M can be written as $M(z_\alpha, z_\beta)$. If z_α and z_β are uncorrelated, removing one from the mapping will reduce the information provided by M. That is, $M(z_\alpha, \emptyset) \neq M(z_\alpha, z_\beta)$ and $M(\emptyset, z_\beta) \neq M(z_\alpha, z_\beta)$.

   Therefore, given transformations T_1 for α and T_2 for domain shifts, and an optimization objective to minimize the covariance between the resulting augmentation-based self-supervised features and invariant features are distinct and uncorrelated in the feature space and both contribute unique information for mapping from feature space to label space.

Hospital-specific representations: "least general or specific features" –

To extract the most specific features, similar to [27], $\gamma(I; \theta_\gamma)$ is used for feature extraction followed by the Δ_γ domain classifier that is trained in a supervised manner using cross-entropy loss to predict the domain/hospital label:

$$\mathcal{L}_{\text{s}} \coloneqq -\mathbb{E}_{(I,.,h)\sim\mathcal{I}} h \log \Delta_\gamma(z^I_\gamma; \theta_{\Delta_\gamma}). \tag{6.4}$$

Disentanglement loss between pairs of extracted features

To prevent redundancy and ensure diversity in our feature extractors, I need to disentangle their resulting representations from each other. This can be achieved by zeroing the covariance matrix between pairs of random vectors, such as z_a and z_b. A zero-covariance matrix indicates that the variables are independent and have no correlation or effect on each other. To enforce this disentanglement, I define pairwise covariance loss functions between each pair of α, β, and γ feature extractors' representations as

$$\mathcal{L}_{\alpha\beta} \coloneqq -\mathbb{E}_{I\sim\mathcal{I}}\big[||\operatorname{Cov}(z^I_\alpha, z^I_\beta)||_2\big], \tag{6.5}$$

$$\mathcal{L}_{\alpha\gamma} \coloneqq -\mathbb{E}_{I\sim\mathcal{I}}\big[||\operatorname{Cov}(z^I_\alpha, z^I_\gamma)||_2\big], \tag{6.6}$$

$$\mathcal{L}_{\beta\gamma} \coloneqq -\mathbb{E}_{I\sim\mathcal{I}}\big[||\operatorname{Cov}(z^I_\beta, z^I_\gamma)||_2\big]. \tag{6.7}$$

Classification loss using aggregation of extracted features –

The goal of the classifier $\Delta_c(z^I_\alpha \oplus z^I_\beta \oplus z^I_\gamma; \theta_c) : z^I_\alpha \oplus z^I_\beta \oplus z^I_\gamma \to \overline{1 : N_c}$ is to classify the images according to their hard class labels using a concatenation of all the extracted features:

$$\mathcal{L}_{\text{c}} \coloneqq -\mathbb{E}_{(I,y,.)\sim\mathcal{I}}\left[y \log \Delta_c(z^I_\alpha \oplus z^I_\beta \oplus z^I_\gamma; \theta_c)\right], \tag{6.8}$$

where y is the target class label for image I.

Inference and training in Phase I –

Using all loss functions, the feature extractors and modules are updated via

$$\mathcal{L}_{\text{total}} := a_1\mathcal{L}_{\text{SSL}} + a_2\mathcal{L}_i + a_3\mathcal{L}_s + a_4\mathcal{L}_{\alpha\beta} + a_5\mathcal{L}_{\alpha\gamma} + a_6\mathcal{L}_{\beta\gamma} + a_7\mathcal{L}_c, \tag{6.9}$$

where a_i coefficients are selected as balanced parameters between loss functions and all were set to 1.0 in our experiments as the loss values were in the same range. Through backward the total loss, i.e., $\mathcal{L}_{\text{total}}$, the updated encoders, i.e., $\alpha'(I;\theta_{\alpha'})$, $\beta'(I;\theta_{\beta'})$, $\gamma'(I;\theta_{\gamma'})$, and classifiers $\Delta'_\beta(I;\theta_{\Delta'_\beta})$, $\Delta'_\gamma(I;\theta_{\Delta'_\gamma})$ and $\Delta'_c(I;\theta_{\Delta'_c})$ are obtained.

6.2.2 Phase II: Meta-learning for generalization improvement

To adapt the domain-specific representation z_γ to the target domain using information from source domains, I use the same meta-learning framework as [27]. The α' and β' feature extractors remain frozen while the γ' feature extractor and Δ'_c classifier are updated. I aim to update $\omega = \theta_{\gamma'} \circ \theta_{\Delta'_c}$ through meta-learning by dividing each hospital data $\mathcal{H}_k$ into disjoint meta-train $\mathcal{H}_k^{tr}$ and meta-test $\mathcal{H}_k^{te}$ sets and the objective is to

$$\min_{\omega} \quad \mathcal{L}_c^{\text{meta}} \coloneqq f(\omega - \nabla f(\omega, \mathcal{H}_k^{tr}), \mathcal{H}_k^{te}), \tag{6.10}$$

where

$$f(\omega = \theta_{\gamma'} \circ \theta_{\Delta'_c}, \mathcal{H}_k) = -\mathbb{E}_{(I_k,y_k,k)\sim\mathcal{H}_k}\left[y_k \log \Delta'_c(z_{\alpha'}^{I_k} \oplus z_{\beta'}^{I_k} \oplus \gamma'(I_k, \theta_{\gamma'}); \theta_{\Delta'_c})\right], \tag{6.11}$$

where y_k and k are the target class label and hospital label, respectively, for image I_k.

6.3 Experiments and Results

The study evaluates the effectiveness of the proposed method, ALFA, against mDSDI [27], HA [208], and ERM through a leave-one-*domain/hospital*-out evaluation using data from multiple hospitals/domains. The evaluation includes reporting "accuracy" for the target (hold-out) domain/hospital, as well as "Area Under the ROC curve (AUROC)" and "recall" metrics for RCC subtyping, which is important for cancer diagnosis tasks.

Table 6.1: Results on PACS dataset

		Accuracy (%)		
Target	Source	ERM	mDSDI [27]	ALFA(mine)
Photo	{A,C,S}	91.98	90.06	95.15
Art	{P,C,S}	76.85	76.27	83.10
Cartoon	{P,A,S}	74.87	76.20	78.71
Sketch	{P,A,C}	76.76	78.85	78.41
Average		80.11 ± 6.76	80.34 ± 5.60	83.75 ± 6.72

6.3.1 Datasets

– **PACS:** P(hoto), A(rt), C(artoon), S(ketch) is a benchmark for DG on natural images. This benchmark includes four domains (Photo, Sketch, Cartoon, Painting), and 7 common categories 'dog', 'elephant', 'giraffe'. 'guitar', 'horse', 'house', and 'person' with a total 9991 images.

– **RCC subtyping dataset from TCGA:** The RCC dataset [208, 92], already introduced in Chapter 5, comprises patches of various RCC subtypes collected from five different hospitals. Due to the absence of certain subtypes, two of the hospitals' data have been merged. The dataset comprises 4 image repositories: (1) HMD, (2) MSKCC, (3) IGC, and (4) NCI from TCGA. The dataset contains $\approx 70k$ patches of size 224×224.

6.3.2 Experimental Setup

The backbone of all feature extractors was the ResNet18 [84], pre-trained on ImageNet [122], with all of its batch normalization layers frozen as per the guidelines given in Seo et al. [197]. All features were embedded to a size of 512. The Adam optimizer was employed with an initial learning rate of 5e-5. A batch size of 32 was established and set the maximum number of iterations to 3000.

6.3.3 Results

– **Losses convergence:** During the training, it was found that γ feature extractor was dominating over other feature extractors and potentially causing a dampening effect on their contributions to the model's overall performance. Hence, I added layer normalization [5] whenever the representations are concatenated to address this issue. With this modification, all losses converged almost simultaneously according to Fig. 6.3.

– **Low-dimensional Embedding Visualization:** In accordance with the best practices suggested in the original UMAP paper [153], I applied PCA [251] to obtain the first 50 principal components, followed by UMAP [153] for further dimensionality reduction to 2.

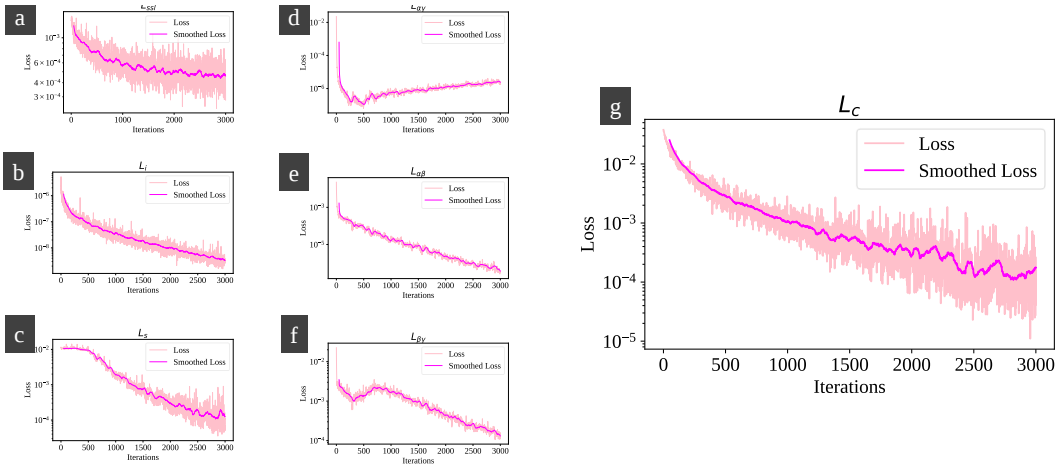


Figure 6.3: The behavior of ALFA's loss functions during the training when the hold-out set was Art on PACS. (a) $\mathcal{L}_{\text{SSL}}$, (b) $\mathcal{L}_i$, (c) $\mathcal{L}_s$, (d) $\mathcal{L}_{\alpha\gamma}$, (e) $\mathcal{L}_{\alpha\beta}$, (f) $\mathcal{L}_{\beta\gamma}$, (g) $\mathcal{L}_c$

According to Figs. 6.5 and 6.4, ALFA's domain-invariant approach yields a more powerful discriminatory representation for different RCC subtypes or different categories on PACS, compared to mDSDI [27], for domain-specific representations. In other words, ALFA's domain-invariant encoder learns some features that are also learned by mDSDI's domain-specific encoder. ALFA's SSL representation provides useful representation, which seems even better than mDSDI's domain-invariant features according to these figures.

PACS dataset classification task

The accuracy of mDSDI [27] and ALFA applied on PACS have been reported in Table 6.1. It can be seen in Table 6.1, except for the 'Sketch' with a high semantic shift in comparison to the rest of the target domains, ALFA outperforms mDSDI with an average accuracy of $83.75 \pm 6.72\%$ compared to mDSDI's average accuracy of $80.34 \pm 5.60\%$. According to this, ALFA cannot only be effective for generalization to unseen hospitals but it can also be effective for DG tasks for natural images.

RCC subtyping classification task

The accuracy of mDSDI [27], HA approach introduced in the previous chapter, and ALFA applied on RCC subtyping task has been reported in Table 6.3.

In the context of HMD, ALFA exceeds the performance of both ERM and mDSDI, achieving an accuracy of 65.52% as opposed to 72.49% and 51.72% respectively. However, it falls short when compared to the HA method, which reaches 75.29% accuracy. This outcome may indicate that, considering HMD encompasses two distinct data sources, methods

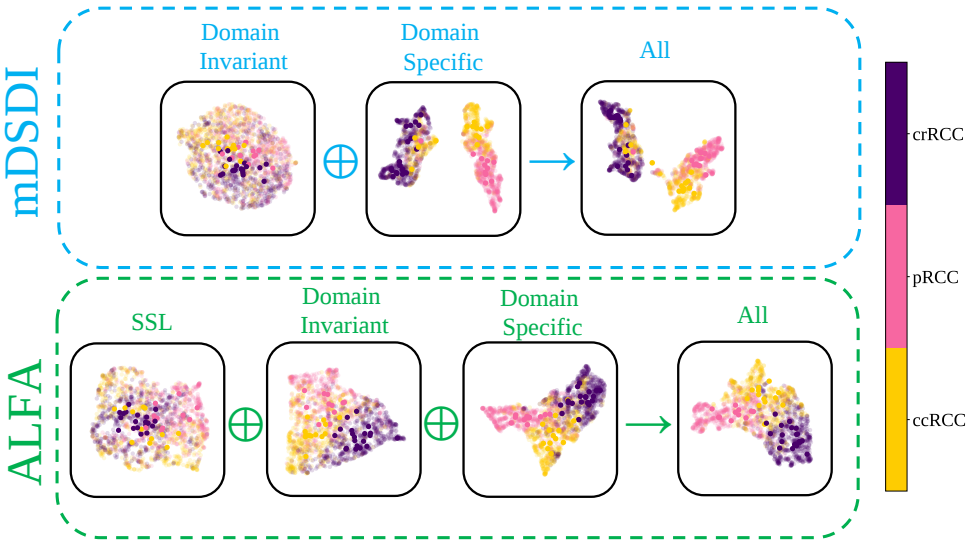


Figure 6.4: 2D feature embeddings for the feature extractors in mDSDI [27] versus in ALFA: (target hospital: 'NCI'). 'All' is the concatenation of domain-specific and domain-invariant representations for the mDSDI [27] (up), and SSL, domain-invariant, and domain-specific representations for ALFA (bottom). Opaque-shaded scatters are WSIs representations obtained by averaging on patches' representations (transparent-shaded).

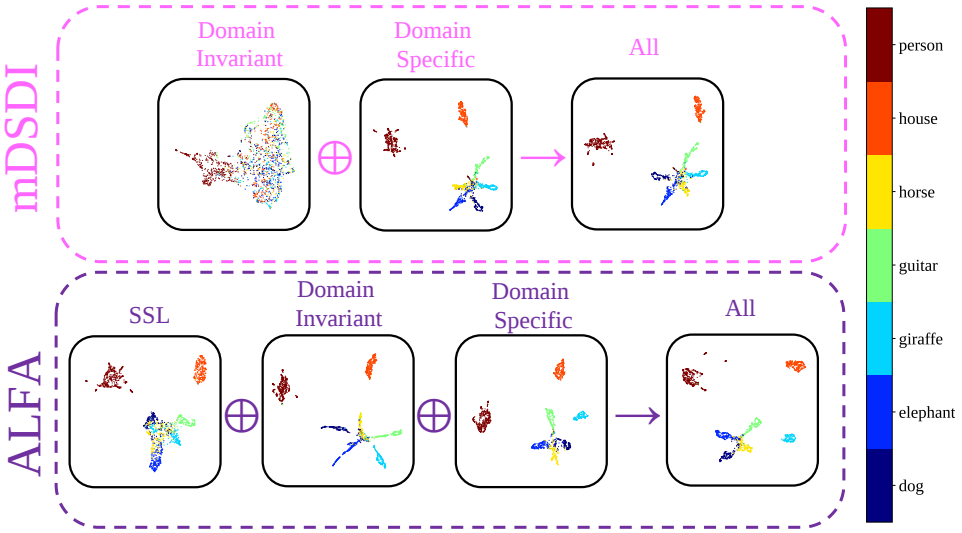


Figure 6.5: 2D feature embedding for the feature extractors in mDSDI (upper row) versus in ALFA (bottom row), target domain:'Photo' on PACS.

that focus on extracting invariant features are more effective. In this regard, HA, which specifically targets hospital-invariant features, outperforms both ALFA and mDSDI.

For the IGC dataset, both ALFA and mDSDI attain the same accuracy of 86.21%, outperforming ERM and HA, which scored 75.86% and 70.42%, respectively. Given that both

ALFA and mDSDI leverage a combination of invariant and hospital-specific features, it's reasonable to speculate that the inclusion of specific features has enhanced their overall performance in this context. In essence, the unique features sourced from the different hospitals appear to have provided valuable information that contributed to improved performance on the IGC dataset.

At the NCI, ALFA again leads with an accuracy of 86.36%, outperforming ERM (81.82%), mDSDI (72.73%), and HA (83.38%).

In the case of MSKCC, ALFA's performance is notably lower with an accuracy of 84.69%, when compared to both ERM (86.73%) and HA (88.19%). However, ALFA does surpass mDSDI, which has an accuracy of 85.71%. This result might suggest that invariant features are more conducive to effective generalization in this context than specific features. Therefore, both mDSDI and ALFA, which rely on a combination of site-specific and site-invariant features, may not perform as well as HA or even ERM.

According to this table, ALFA outperforms mDSDI, HA, and ERM with average accuracy, AUROC, and recall of $80.69 \pm 8.61\%$, $94.90 \pm 2.66\%$, and $81.62 \pm 8.50\%$, respectively. ALFA performed similarly to mDSDI and ERM in terms of AUROC and accuracy, but slightly better overall. However, ALFA significantly outperformed other methods in terms of recall metric, especially for the "NCI", and "IGC" target hospitals.

Ablation study

To further investigate ALFA, an ablation study was conducted indicating that all components are effective in achieving generalization according to Table 6.2. In this table, the feature extractors under consideration are α, β, and γ, and their activation or deactivation states are indicated in blue (active) or gray (inactive). The performance is evaluated across four categories: Photo, Art, Cartoon, and Sketch, and an average score is calculated. When α is active and the other two feature extractors are deactivated, the average accuracy is the lowest, 48.63 ± 14.98. However, when either β or γ is active with the others deactivated, there's a significant increase in the average accuracy (around 78.88 ± 6.99 and 78.82 ± 9.11 respectively). The highest average accuracies are achieved when multiple features are activated together, with the complete ALFA method, where all feature extractors are active, achieving the highest average accuracy of 84.09 ± 7.06.

6.4 Conclusion

An innovative approach was proposed coined ALFA leverages multiple feature extractors to disentangle features at different levels of abstraction. The effectiveness of ALFA was demonstrated on various benchmarks, including PACS, a widely used DG benchmark, as well as RCC subtype classification from TCGA database. ALFA outperformed the state-of-the-art mDSDI [27] approach on PACS, highlighting its superior performance.

Table 6.2: Ablation on PACS: Active feature extractor(s) is/are blue, Deactivated one(s): gray

	Photo	Art	Cartoon	Sketch	Average
α, β, γ	69.70	53.17	44.28	27.38	48.63 ± 14.98
α, β, γ	91.13	73.33	75.68	75.38	78.88 ± 6.99
α, β, γ	94.43	77.34	72.05	71.46	78.82 ± 9.11
α, β, γ	90.11	75.48	76.23	75.99	79.45 ± 6.03
α, β, γ	94.73	78.61	73.93	72.02	79.82 ± 8.75
α, β, γ	95.38	82.03	78.37	78.67	83.61 ± 6.80
ALFA	96.15	83.10	78.71	78.41	84.09 ± 7.06

Furthermore, ALFA demonstrated excellent generalization capabilities to unseen hospitals in comparison to mDSDI and ERM. In this study, I used a simple augmentation-based self-supervision technique. However, there is a lot of room to improve the proposed approach by integrating more effective SSL features, especially with ongoing research in self-supervision. In addition, to achieve hospital-invariant features a new form of domain alignment loss function called "*soft class-domain alignment*" loss was proposed. It effectively extracts domain/hospital-invariant features that align the data of two different sources with soft class labels as a reference distribution, making it a valuable addition to my proposed design.

Table 6.3: Results on RCC subtyping task

		Accuracy (%)				AUROC (%)				Recall (%)			
Target	Source	ERM	mDSDI [27]	HA [208]	ALFA	ERM	mDSDI [27]	HA [208]	ALFA	ERM	mDSDI [27]	HA [208]	ALFA
IGC	{NCI, MSKCC, HMD}	75.86	86.20	70.42	86.21	93.23	95.78	88.36	95.33	57.14	82.88	62.38	85.39
NCI	{IGC, MSKCC, HMD}	81.82	72.73	83.38	86.36	96.49	94.46	97.32	97.83	83.08	71.46	85.48	86.41
MSKCC	{IGC, NCI, HMD}	86.73	85.71	88.19	84.69	95.91	95.89	96.47	95.99	82.99	87.05	85.32	87.99
HMD	{IGC, NCI, MSKCC}	72.49	51.72	75.29	65.52	85.38	88.37	90.16	90.48	72.96	51.85	78.42	66.67
Average		79.22 ±5.36	74.09 ±13.72	79.32 ±6.49	80.69 ±8.61	92.75 ±4.34	93.62 ±3.02	93.08 ±3.34	94.90 ±2.66	74.07 ±10.38	73.31 ±13.36	77.90 ±8.44	81.62 ±8.50

Chapter 7

Summary and Future Directions

The journey of this thesis has involved a structured exploration of two crucial forms of generalization in the domain of histopathology. The focus on magnification generalization and OOD generalization towards different hospitals, as well as the proposition of a HA learning regime and a comprehensive method, i.e., ALFA, have shaped the main contours of this study.

The exploration of magnification generalization in Chapter 3 recognized the variability in histopathological images due to distinct magnification levels. The proposed method sought to enhance the robustness and reliability of the model by learning invariant features at these varying levels. This led to the development of a model that could reliably identify and learn consistent features across different settings and equipment.

In Chapter 4, the attention was shifted to OOD generalization, with a particular emphasis on unseen hospital data. This investigation was built on the premise that pre-existing models could house valuable knowledge, capable of equipping new models to adapt to diverse data scenarios. The aim was to ensure that these models could operate effectively in different hospital environments.

Chapter 5 introduced the concept of HA, highlighting the invariant characteristics in hospitals. The goal was to develop a learning model that maintained its stability in performance in various hospital settings, ensuring its utility and reliability in a variety of clinical settings.

Lastly, in Chapter 6, the thesis presented a comprehensive method that considered not just invariant features across hospitals, but an expanded set of features extractable from the input images. This multifaceted approach was designed to maximize the generalization potential of the model, thereby contributing to the progress of diagnostic accuracy in histopathology.

The implications of these findings are profound, pointing to a path toward a better generalization of diagnostic precision in histopathology by harnessing the power of DG. Even though, there are some areas that have not been explored and can be framed as the future directions of the current study.

7.1 Future Directions

Looking ahead, several promising avenues for future research are worth exploring.

7.1.1 Generalization to Unseen Classes in Histopahtology

One of the directions of future work for this thesis involves the progression to generalization to "unseen classes" in histopathology [259, 199]. This important concept refers to the ability of a machine learning model to accurately predict or classify new classes (e.g., different unseen cancer sub-types) that it has not been explicitly trained on, which poses a significant challenge in the medical field due to the diverse and evolving nature of diseases.

In the context of histopathology, the generalization to unseen classes can be paramount for the diagnosis and study of rare or novel diseases, enhancing patient outcomes by enabling more timely and accurate diagnoses. Moreover, it could revolutionize telemedicine by empowering machines to adapt to unseen conditions without continuous retraining, thereby saving crucial time and resources [22].

This objective will be pursued through the application of few-shot learning and meta-learning approaches [248, 259]. Few-shot learning is a methodology in which the ML model is designed to make accurate predictions with only a few training examples for each new class. This approach is especially relevant in histopathology, where certain classes of disease may have limited sample availability.

Furthermore, the exploration of MAML (or "*learning to learn*") [62, 114] will be instrumental in this journey. MAML learn the strategy of learning itself, aiming to swiftly adapt to new tasks with minimal data. This characteristic is highly beneficial in a field like histopathology where novel classes can frequently emerge.

The evaluation of these methodologies will be performed through the N-way K-shot classification. This type of evaluation involves classifying K examples each from N classes, and it is an established way to measure the performance of models in a few-shot learning setup.

Ultimately, the move towards generalization to unseen classes in histopathology aligns with the overarching goal of this research: improving the adaptability and performance of machine learning models in real-world medical scenarios, thereby contributing to the advancement of AI in healthcare.

7.1.2 Multi-Instance Learning instead of Pure Weakly-Supervision

Another exciting future direction for this thesis involves leveraging attention-based Multiple-Instance Learning (MIL) [99] to mitigate the bias introduced by weakly-supervised learning and improve the assignment of patch labels to their parent WSIs [146].

In the context of digital pathology, WSIs often consist of billions of pixels, making them too large to process in their entirety. As a solution, these WSIs are divided into smaller and more manageable sections known as patches. However, there is a challenge in assigning labels to these patches based on the labels of their parent WSIs, as this can result in inaccuracies due to the inherent heterogeneity within the WSIs. This label assignment process is an example of weakly-supervised learning, where only coarse or imprecise labels are available.

Attention-based MIL presents a promising solution to this issue. In the attention-based MIL framework, training examples are organized into bags, and a label is assigned to each bag rather than individual instances. In the case of digital pathology, a bag could represent a WSI, and the instances would be the patches derived from that WSI [146].

The key advantage of attention-based MIL is that it can handle ambiguity in instance-level labels, which aligns perfectly with the nature of histopathological data. By treating each WSI as a bag, we can make the learning process more robust against potential mislabeling introduced by weak supervision.

Moreover, attention-based MIL can be utilized to automatically identify and learn from the most informative patches within WSI. This allows the model to focus on the most relevant areas of the image, enhancing the accuracy of disease classification and prognosis prediction.

7.1.3 Distribution Shift Quantification

Yet another promising avenue for future research in this thesis could be the quantification of the distribution shift [157, 226] in histopathological images [217, 163]. Recognizing and addressing distribution shifts is critical to ensure the robust performance of ML models, especially in the medical domain where data can vary significantly due to factors like differing equipment or procedures across hospitals, or inherent biological variability amongst patients.

While methods such as Proxy-A Distance (PAD) [17] provide a useful tool for estimating the shift between distributions, they do have limitations. One of the primary drawbacks is that PAD is dependent on the classifier being trained, which could introduce a bias in the estimation. Additionally, it might not provide an accurate measure in scenarios where the shift is subtle or complex in nature.

Therefore, the development of a more universal, classifier-agnostic metric to recognize and quantify the distribution shift would be invaluable. This metric should be able to accurately capture the degree of shift regardless of the nature of the classifier used. It should also be sensitive to even minor shifts, which can be crucial in a field like histopathology where small changes in cellular or tissue structure can signal significant pathological changes.

Quantification of distribution shift could lead to more reliable predictions by allowing better calibration of models to new data. Furthermore, it would provide valuable insights

into the robustness of the models and guide the creation of strategies to tackle distribution shifts, such as domain adaptation methods or model retraining schedules.

In essence, advancing towards a more robust and comprehensive quantification of distribution shifts in histopathology data can significantly enhance the model's reliability and performance, thereby leading to more accurate diagnoses and, ultimately, better patient outcomes.

7.1.4 XAI for Explainability of Biases and Shifts

A promising avenue for future research within the scope of this thesis involves the incorporation of XAI methods [90] to interpret biases and shifts within the realm of histopathology and digital pathology [49]. XAI, by providing understandable and interpretable models, will help address one of the significant challenges that currently limit the broader application of ML in the healthcare sector - the lack of transparency in model predictions [80].

The application of XAI could substantially improve the understanding of biases that could be introduced due to weak supervision or dataset-specific characteristics. It can elucidate the specific features on which the model is relying for its predictions, whether these are meaningful pathologic indicators or merely artifacts of data collection and preprocessing. This would not only allow for the identification of these biases but also facilitate their correction, thereby improving the overall accuracy and reliability of the models [80].

Moreover, XAI can also be instrumental in interpreting and managing distribution shifts. By exposing how the model's decisions change with varying data distributions, it would be possible to identify if and when the model is struggling due to a distribution shift. Such insights could guide the development of robust strategies to mitigate the impacts of distribution shifts, including DG or DA techniques or model recalibration schedules.

Furthermore, the interpretability provided by XAI is invaluable in establishing trust and facilitating the acceptance of AI models among healthcare professionals. Understanding why a model makes a particular prediction fosters confidence in its recommendations, which is critical in a high-stakes field like healthcare.

In conclusion, integrating XAI for the explainability of biases and shifts in histopathology could significantly contribute towards enhancing model performance, understanding, and acceptance, thereby promoting the efficient application of AI in the field of digital pathology.

7.1.5 Improvement on ALFA

In this section, I present several prospective paths for the advancement of the ALFA framework, which was first introduced in this thesis in Chapter 6. These potential future directions include:

a) As a potential future direction, I propose the integration of more advanced self-supervision models [32] into the ALFA. Such enhanced self-supervision techniques can potentially offer more refined segregation in the feature space, contributing to a higher level of abstraction. This refined abstraction is anticipated to further enhance the robustness of the model, helping it generalize better across varied scenarios and handle more complex tasks. The study of how different self-supervision techniques could improve the performance of the ALFA framework, and the exploration of novel techniques designed specifically for this purpose, can provide intriguing avenues for future research in histopathology. The overall aim is to keep refining and optimizing the ALFA framework for the betterment of diagnostic accuracy in the field of histopathology.

b) In a future line of investigation for the ALFA framework, I plan to implement mutual information methodologies as a potential replacement for the covariance operations currently used for feature disentanglement. This shift is motivated by the inherent limitations of covariance in only capturing linear relationships [82], which may not encapsulate the complexity of interactions in high-dimensional feature spaces typically encountered in histopathology images. Mutual information, on the other hand, is capable of capturing both linear and non-linear dependencies between variables [24], thus providing a more comprehensive measure of correlation. By integrating mutual information into the ALFA framework, we could potentially attain a more robust and complete disentanglement of features. This could lead to a richer, more accurate representation of histopathological images, thereby pushing the boundaries of diagnostic precision in histopathology further. This proposed enhancement marks an exciting direction for future research, which could significantly advance our understanding and utilization of the ALFA framework.

References

[1] Balázs Acs, Mattias Rantalainen, and Johan Hartman. Artificial intelligence as the next step towards precision pathology. *Journal of internal medicine*, 288(1):62–81, 2020.

[2] Roee Aharoni and Yoav Goldberg. Split and rephrase: Better evaluation and a stronger baseline. *arXiv preprint arXiv:1805.01035*, 2018.

[3] Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8:1–74, 2021.

[4] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[5] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[6] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. Neural codes for image retrieval. In *European conference on computer vision*, pages 584–599. Springer, 2014.

[7] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *Advances in Neural Information Processing Systems*, 31:998–1008, 2018.

[8] Elisa V Bandera, Gertraud Maskarinec, Isabelle Romieu, and Esther M John. Racial and ethnic disparities in the impact of obesity on breast cancer risk and survival: a global perspective. *Advances in Nutrition*, 6(6):803–819, 2015.

[9] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018.

[10] Yaniv Bar, Idit Diamant, Lior Wolf, Sivan Lieberman, Eli Konen, and Hayit Greenspan. Chest pathology detection using deep learning with non-medical training. In *2015 IEEE 12th international symposium on biomedical imaging (ISBI)*, pages 294–297. IEEE, 2015.

[11] Neslihan Bayramoglu and Janne Heikkilä. Transfer learning for cell nuclei classification in histopathology images. In *European Conference on Computer Vision*, pages 532–539. Springer, 2016.

[12] Neslihan Bayramoglu, Juho Kannala, and Janne Heikkilä. Deep learning for magnification independent breast cancer histopathology image classification. In *2016 23rd International Conference on Pattern Recognition*, pages 2440–2445. IEEE, 2016.

[13] Andrew H Beck, Ankur R Sangoi, Samuel Leung, Robert J Marinelli, Torsten O Nielsen, Marc J Van De Vijver, Robert B West, Matt Van De Rijn, and Daphne Koller. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Science translational medicine*, 3(108):108ra113–108ra113, 2011.

[14] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.

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

[16] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010.

[17] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006.

[18] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[19] Yoshua Bengio, Aaron C Courville, and Pascal Vincent. Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR, abs/1206.5538*, 1:2012, 2012.

[20] Aïcha BenTaieb and Ghassan Hamarneh. Topology aware fully convolutional networks for histology gland segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 460–468. Springer, 2016.

[21] Christof A Bertram, Marc Aubreville, Christian Marzahl, Andreas Maier, and Robert Klopfleisch. A large-scale dataset for mitotic figure assessment on whole slide images of canine cutaneous mast cell tumor. *Scientific data*, 6(1):1–9, 2019.

[22] Samar Betmouni. Diagnostic digital pathology implementation: Learning from the digital health experience. *Digital Health*, 7:20552076211020240, 2021.

[23] Hakan Bilen and Andrea Vedaldi. Universal representations: The missing link between faces, text, planktons, and cat breeds. *arXiv preprint arXiv:1701.07275*, 2017.

[24] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[25] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in neural information processing systems*, pages 343–351, 2016.

[26] Jane Bromley, James W Bentz, Léon Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(04):669–688, 1993.

[27] Manh-Ha Bui, Toan Tran, Anh Tran, and Dinh Phung. Exploiting domain-specific features to enhance domain generalization. *Advances in Neural Information Processing Systems*, 34:21189–21201, 2021.

[28] Jiatong Cai, Chenglu Zhu, Can Cui, Honglin Li, Tong Wu, Shichuan Zhang, and Lin Yang. Generalizing nucleus recognition model in multi-source ki67 immunohistochemistry stained images via domain-specific pruning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 277–287. Springer, 2021.

[29] 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.

[30] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Partial transfer learning with selective adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2724–2732, 2018.

[31] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

[32] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[33] Jia-Ren Chang, Min-Sheng Wu, Wei-Hsiang Yu, Chi-Chung Chen, Cheng-Kung Yang, Yen-Yu Lin, and Chao-Yuan Yeh. Stain mix-up: Unsupervised domain generalization for histopathology images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 117–126. Springer, 2021.

[34] K Chang, CJ Creighton, C Davis, L Donehower, et al. The cancer genome atlas pan-cancer analysis project. *Nat Genet*, 45(10):1113–1120, 2013.

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

[36] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[37] Zhenzhou Chen, Dmitriy Shin, Shanyan Chen, Kovalenko Mikhail, Orr Hadass, Brittany N Tomlison, Dmitry Korkin, Chi-Ren Shyu, Jiankun Cui, Douglas C Anthony, et al. Histological quantitation of brain injury using whole slide imaging: a pilot validation study in mice. *PLoS One*, 9(3):e92133, 2014.

[38] Jun Cheng, Jie Zhang, Yatong Han, Xusheng Wang, Xiufen Ye, Yuebo Meng, Anil Parwani, Zhi Han, Qianjin Feng, and Kun Huang. Integrative analysis of histopathological images and genomic data predicts clear cell renal cell carcinoma prognosis. *Cancer research*, 77(21):e91–e100, 2017.

[39] John C Cheville, Christine M Lohse, Horst Zincke, Amy L Weaver, and Michael L Blute. Comparisons of outcome and prognostic features among histologic subtypes of renal cell carcinoma. *The American journal of surgical pathology*, 27(5):612–624, 2003.

[40] Deborah L Commins, Roscoe D Atkinson, and Margaret E Burnett. Review of meningioma histopathology. *Neurosurgical focus*, 23(4):E3, 2007.

[41] 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*, 24(10):1559–1567, 2018.

[42] Kausik Das, Sri Phani Krishna Karri, Abhijit Guha Roy, Jyotirmoy Chatterjee, and Debdoot Sheet. Classifying histopathology whole-slides using fusion of decisions from deep convolutional network on a collection of random multi-views at multi-magnification. In *2017 IEEE 14th International Symposium on Biomedical Imaging*, pages 1024–1027. IEEE, 2017.

[43] Hal Daumé III. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*, 2009.

[44] Taher Dehkharghanian, Azam Asilian Bidgoli, Abtin Riasatian, Pooria Mazaheri, Clinton JV Campbell, Liron Pantanowitz, HR Tizhoosh, and Shahryar Rahnamayan. Biased data, biased ai: deep networks predict the acquisition site of tcga images. *Diagnostic pathology*, 18(1):1–12, 2023.

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

[46] Alex Diaz-Papkovich, Luke Anderson-Trocmé, Chief Ben-Eghan, and Simon Gravel. Umap reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS genetics*, 15(11):e1008432, 2019.

[47] Neofytos Dimitriou, Ognjen Arandjelović, and Peter D Caie. Deep learning for whole slide image analysis: an overview. *Frontiers in medicine*, 6:264, 2019.

[48] Thanh-Toan Do, Toan Tran, Ian Reid, Vijay Kumar, Tuan Hoang, and Gustavo Carneiro. A theoretically sound upper bound on the triplet loss for improving the efficiency of deep distance metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10404–10413, 2019.

[49] James M Dolezal, Rachelle Wolk, Hanna M Hieromnimon, Frederick M Howard, Andrew Srisuwananukorn, Dmitry Karpeyev, Siddhi Ramesh, Sara Kochanny, Jung Woo Kwon, Meghana Agni, et al. Deep learning generates synthetic cancer histology for explainability and education. *NPJ Precision Oncology*, 7(1):49, 2023.

[50] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655, 2014.

[51] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[52] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *Advances in Neural Information Processing Systems*, pages 6450–6461, 2019.

[53] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM Transactions on graphics (TOG)*, 31(4):1–10, 2012.

[54] 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.

[55] Freja Fagerblom. Model-agnostic meta-learning for digital pathology. Master's thesis, Linköping University, Department of Electrical Engineering, 2020.

[56] Min Fang, Yong Guo, Xiaosong Zhang, and Xiao Li. Multi-source transfer learning based on label shared subspace. *Pattern Recognition Letters*, 51:101–106, 2015.

[57] Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R. Arabnia. A Brief Review of Domain Adaptation. 2020.

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

[59] Kevin Faust, Sudarshan Bala, Randy van Ommeren, Alessia Portante, Raniah Al Qawahmed, Ugljesa Djuric, and Phedias Diamandis. Intelligent feature engineering and ontological mapping of brain tumour histomorphologies by deep learning. *Nature Machine Intelligence*, 1(7):316–321, 2019.

[60] Kevin Faust, Adil Roohi, Alberto J Leon, Emeline Leroux, Anglin Dent, Andrew J Evans, Trevor J Pugh, Sangeetha N Kalimuthu, Ugljesa Djuric, and Phedias Diamandis. Unsupervised resolution of histomorphologic heterogeneity in renal cell carcinoma using a brain tumor–educated neural network. *JCO Clinical Cancer Informatics*, 4:811–821, 2020.

[61] Kevin Faust, Adil Roohi, Alberto J. Leon, Emeline Leroux, Anglin Dent, Andrew J. Evans, Trevor J. Pugh, Sangeetha N. Kalimuthu, Ugljesa Djuric, and Phedias Diamandis. Unsupervised Resolution of Histomorphologic Heterogeneity in Renal Cell Carcinoma Using a Brain Tumor–Educated Neural Network. *JCO Clinical Cancer Informatics*, (4):811–821, 2020.

[62] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *The 34th International Conference on Machine Learning*, 2017.

[63] Kunihiko Fukushima. Cognitron: A self-organizing multilayered neural network. *Biological cybernetics*, 20(3):121–136, 1975.

[64] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[65] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[66] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[67] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[68] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. *Advances in neural information processing systems*, 31, 2018.

[69] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, pages 2551–2559, 2015.

[70] Benyamin Ghojogh, Milad Sikaroudi, Sobhan Shafiei, Hamid R Tizhoosh, Fakhri Karray, and Mark Crowley. Fisher discriminant triplet and contrastive losses for training siamese networks. In *2020 international joint conference on neural networks (IJCNN)*, pages 1–7. IEEE, 2020.

[71] Benyamin Ghojogh, Milad Sikaroudi, Hamid R Tizhoosh, Fakhri Karray, and Mark Crowley. Weighted fisher discriminant analysis in the input and feature spaces. In *International Conference on Image Analysis and Recognition*, pages 3–15. Springer, 2020.

[72] Jacob Gildenblat and Eldad Klaiman. Self-supervised similarity learning for digital pathology. In *MICCAI 2019 COMPAY Workshops*, 2019.

[73] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[74] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[75] Michel Goossens, Frank Mittelbach, and Alexander Samarin. *The LaTeX Companion*. Addison-Wesley, Reading, Massachusetts, 1994.

[76] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

[77] Metin N Gurcan, Laura E Boucheron, Ali Can, Anant Madabhushi, Nasir M Rajpoot, and Bulent Yener. Histopathological image analysis: A review. *IEEE reviews in biomedical engineering*, 2:147–171, 2009.

[78] David A Gutman, Jake Cobb, Dhananjaya Somanna, Yuna Park, Fusheng Wang, Tahsin Kurc, Joel H Saltz, Daniel J Brat, Lee AD Cooper, and Jun Kong. Cancer digital slide archive: an informatics resource to support integrated in silico analysis of tcga pathology data. *Journal of the American Medical Informatics Association*, 20(6):1091–1098, 2013.

[79] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[80] Miriam Hägele, Philipp Seegerer, Sebastian Lapuschkin, Michael Bockmayr, Wojciech Samek, Frederick Klauschen, Klaus-Robert Müller, and Alexander Binder. Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. *Scientific reports*, 10(1):1–12, 2020.

[81] Seokmin Han, Sung Il Hwang, and Hak Jong Lee. The classification of renal cancer in 3-phase ct images using a deep learning method. *Journal of digital imaging*, 32(4):638–643, 2019.

[82] Wolfgang Karl Härdle and Léopold Simar. *Applied multivariate statistical analysis*. Springer Nature, 2019.

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

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

[85] Jeff Heaton. Ian goodfellow, yoshua bengio, and aaron courville: Deep learning: The mit press, 2016, 800 pp, isbn: 0262035618. *Genetic Programming and Evolvable Machines*, 19(1-2):305–307, 2018.

[86] Thomas Hellström, Virginia Dignum, and Suna Bensch. Bias in machine learning–what is it good for? *arXiv preprint arXiv:2004.00686*, 2020.

[87] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021.

[88] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer, 2015.

[89] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018.

[90] Andreas Holzinger, Anna Saranti, Christoph Molnar, Przemyslaw Biecek, and Wojciech Samek. Explainable ai methods-a brief overview. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, pages 13–38. Springer, 2020.

[91] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *arXiv preprint arXiv:2004.05439*, 2020.

[92] Milad Hosseini, S Maryamand Sikaroudi, Morteza Babaie, and HR Tizhoosh. Proportionally fair hospital collaborations in federated learning of histopathology images. *IEEE Transactions on Medical Imaging*, 2023.

[93] James J Hsieh, Mark P Purdue, Sabina Signoretti, Charles Swanton, Laurence Albiges, Manuela Schmidinger, Daniel Y Heng, James Larkin, and Vincenzo Ficarra. Renal cell carcinoma. *Nature reviews Disease primers*, 3(1):1–19, 2017.

[94] Kyle Hsu, Sergey Levine, and Chelsea Finn. Unsupervised learning via meta-learning. *arXiv preprint arXiv:1810.02334*, 2018.

[95] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. *arXiv preprint arXiv:1711.10125*, 2017.

[96] Gao Huang, Shichen Liu, Laurens Van der Maaten, and Kilian Q Weinberger. Condensenet: An efficient densenet using learned group convolutions. In *Proceedings of*

the IEEE conference on computer vision and pattern recognition, pages 2752–2761, 2018.

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

[98] Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016.

[99] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

[100] National Cancer Institute. The cancer genome atlas. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga, n.d. Accessed: [insert date accessed].

[101] Humayun Irshad, Antoine Veillard, Ludovic Roux, and Daniel Racoceanu. Methods for nuclei detection, segmentation, and classification in digital histopathology: a review—current status and future potential. *IEEE reviews in biomedical engineering*, 7:97–114, 2013.

[102] Takumi Ishikawa, Junko Takahashi, Mai Kasai, Takayuki Shiina, Yuka Iijima, Hiroshi Takemura, Hiroshi Mizoguchi, and Takeshi Kuwata. Support system for pathologists and researchers. *Journal of pathology informatics*, 6, 2015.

[103] Katarzyna Janocha and Wojciech Marian Czarnecki. On loss functions for deep neural networks in classification. *arXiv preprint arXiv:1702.05659*, 2017.

[104] James V Jester, Peter M Andrews, W Matthew Petroll, Michael A Lemp, and H Dwight Cavanagh. In vivo, real-time confocal imaging. *Journal of electron microscopy technique*, 18(1):50–60, 1991.

[105] Oscar Jimenez-del Toro, Sebastian Otálora, Mats Andersson, Kristian Eurén, Martin Hedlund, Mikael Rousson, Henning Müller, and Manfredo Atzori. Analysis of histopathology images: From traditional machine learning to deep learning. In *Biomedical Texture Analysis*, pages 281–314. Elsevier, 2017.

[106] Shivam Kalra, Hamid R Tizhoosh, Sultaan Shah, Charles Choi, Savvas Damaskinos, Amir Safarpoor, Sobhan Shafiei, Morteza Babaie, Phedias Diamandis, Clinton JV Campbell, et al. Pan-cancer diagnostic consensus through searching archival histopathology images using artificial intelligence. *NPJ digital medicine*, 3(1):1–15, 2020.

[107] Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue. *Zenodo10*, 5281, 2018.

[108] Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine*, 16(1), 2019.

[109] Mahmut Kaya and Hasan Şakir Bilge. Deep metric learning: A survey. *Symmetry*, 11(9):1066, 2019.

[110] Siavash Khodadadeh, Ladislau Bölöni, and Mubarak Shah. Unsupervised meta-learning for few-shot image classification. *arXiv preprint arXiv:1811.11819*, 2018.

[111] Siavash Khodadadeh, Ladislau Boloni, and Mubarak Shah. Unsupervised meta-learning for few-shot image classification. *Advances in neural information processing systems*, 32, 2019.

[112] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *European Conference on Computer Vision*, pages 158–171. Springer, 2012.

[113] Brady Kieffer, Morteza Babaie, Shivam Kalra, and Hamid R Tizhoosh. Convolutional neural networks for histopathology image classification: Training vs. using pre-trained networks. In *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2017.

[114] Taesup Kim, Jaesik Yoon, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. Bayesian model-agnostic meta-learning. *arXiv preprint arXiv:1806.03836*, 2018.

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

[116] Donald Knuth. *The TEXbook*. Addison-Wesley, Reading, Massachusetts, 1986.

[117] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, 2015.

[118] Daisuke Komura and Shumpei Ishikawa. Machine learning methods for histopathological image analysis. *Computational and structural biotechnology journal*, 16:34–42, 2018.

[119] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671, 2019.

[120] Wouter M Kouw and Marco Loog. A review of domain adaptation without target labels. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):766–785, 2019.

[121] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

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

[123] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[124] Abhishek Kumar, Avishek Saha, and Hal Daume. Co-regularization based semi-supervised domain adaptation. In *Advances in neural information processing systems*, pages 478–486, 2010.

[125] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022.

[126] Maxime W Lafarge, Josien PW Pluim, Koen AJ Eppenhof, Pim Moeskops, and Mitko Veta. Domain-adversarial neural networks to address the appearance variability of histopathology images. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 83–91. Springer, 2017.

[127] Maxime W. Lafarge, Josien P.W. Pluim, Koen A.J. Eppenhof, and Mitko Veta. Learning Domain-Invariant Representations of Histological Images. *Frontiers in Medicine*, 6:162, 2019.

[128] Leslie Lamport. *LaTeX — A Document Preparation System*. Addison-Wesley, Reading, Massachusetts, second edition, 1994.

[129] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

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

[131] Alona Levy-Jurgenson, Xavier Tekpli, Vessela N Kristensen, and Zohar Yakhini. Spatial transcriptomics inferred from pathology whole-slide images links tumor heterogeneity to survival in breast and lung cancer. *Scientific reports*, 10(1):1–11, 2020.

[132] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Deeper, broader and artier domain generalization. In *International Conference on Computer Vision*, 2017.

[133] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[134] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[135] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 624–639, 2018.

[136] Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[137] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[138] Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob van de Loo, Rob Vogels, et al. 1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset. *GigaScience*, 7(6):giy065, 2018.

[139] Yun Liu, Krishna Gadepalli, Mohammad Norouzi, George E Dahl, Timo Kohlberger, Aleksey Boyko, Subhashini Venugopalan, Aleksei Timofeev, Philip Q Nelson, Greg S Corrado, et al. Detecting cancer metastases on gigapixel pathology images. *arXiv preprint arXiv:1703.02442*, 2017.

[140] Zhenhua Liu, Jizheng Xu, Xiulian Peng, and Ruiqin Xiong. Frequency-domain dynamic pruning for convolutional neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 1051–1061, 2018.

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

[142] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.

[143] Antonio Lopez-Beltran, Jose C Carrasco, Liang Cheng, Marina Scarpelli, Ziya Kirkali, and Rodolfo Montironi. 2009 update on the classification of renal epithelial tumors in adults. *International Journal of Urology*, 16(5):432–443, 2009.

[144] Gavin Low, Guan Huang, Winnie Fu, Zaahir Moloo, and Safwat Girgis. Review of renal cell carcinoma and its common subtypes in radiology. *World journal of radiology*, 8(5):484, 2016.

[145] Hrushikesh Loya, Pranav Poduval, Deepak Anand, Neeraj Kumar, and Amit Sethi. Uncertainty estimation in cancer survival prediction. *arXiv preprint arXiv:2003.08573*, 2020.

[146] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

[147] Siyuan Lu, Zhihai Lu, and Yu-Dong Zhang. Pathological brain detection based on alexnet and transfer learning. *Journal of computational science*, 30:41–47, 2019.

[148] Xu Luo, Longhui Wei, Liangjian Wen, Jinrong Yang, Lingxi Xie, Zenglin Xu, and Qi Tian. Rectifying the shortcut learning of background for few-shot learning. *Advances in Neural Information Processing Systems*, 34:13073–13085, 2021.

[149] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2507–2516, 2019.

[150] Jianzhu Ma, Samson H Fong, Yunan Luo, Christopher J Bakkenist, John Paul Shen, Soufiane Mourragui, Lodewyk FA Wessels, Marc Hafner, Roded Sharan, Jian Peng, et al. Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. *Nature Cancer*, pages 1–12, 2021.

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

[152] Massimiliano Mancini, Samuel Rota Bulo, Barbara Caputo, and Elisa Ricci. Best sources forward: domain generalization through source-specific nets. In *2018 25th IEEE international conference on image processing (ICIP)*, pages 1353–1357. IEEE, 2018.

[153] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[154] L Jeffrey Medeiros, Edward C Jones, Shigeo Aizawa, Hector C Aldape, John C Cheville, Neal S Goldstein, Irina A Lubensky, Jae Ro, Jonathan Shanks, Anna Pacelli, et al. Grading of renal cell carcinoma: Workgroup no. 2. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 80(5):990–991, 1997.

[155] Alfonso Medela, Artzai Picon, Cristina L Saratxaga, Oihana Belar, Virginia Cabezón, Riccardo Cicchi, Roberto Bilbao, and Ben Glover. Few shot learning in histopathological images: reducing the need of labeled data on biological datasets. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1860–1864. IEEE, 2019.

[156] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[157] Timo Milbich, Karsten Roth, Samarth Sinha, Ludwig Schmidt, Marzyeh Ghassemi, and Bjorn Ommer. Characterizing generalization under out-of-distribution shifts in deep metric learning. *Advances in Neural Information Processing Systems*, 34:25006–25018, 2021.

[158] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3523–3542, 2021.

[159] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5715–5725, 2017.

[160] Robert J Motzer, Neil H Bander, and David M Nanus. Renal-cell carcinoma. *New England Journal of Medicine*, 335(12):865–875, 1996.

[161] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18, 2013.

[162] Valdair F Muglia and Adilson Prando. Renal cell carcinoma: histological classification and correlation with imaging findings. *Radiologia brasileira*, 48:166–174, 2015.

[163] Zeeshan Nisar, Jelica Vasiljević, Pierre Gançarski, and Thomas Lampert. Towards measuring domain shift in histopathological stain translation in an unsupervised manner. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2022.

[164] World Health Organization. Cancer. https://www.who.int/news-room/fact-sheets/detail/cancer, 2021.

[165] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

[166] Raju Pal and Mukesh Saraswat. Enhanced bag of features using alexnet and improved biogeography-based optimization for histopathological image analysis. In

2018 eleventh international conference on contemporary computing (IC3), pages 1–6. IEEE, 2018.

[167] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 754–763, 2017.

[168] Liron Pantanowitz. Digital images and the future of digital pathology. *Journal of pathology informatics*, 1, 2010.

[169] Liron Pantanowitz, John H Sinard, Walter H Henricks, Lisa A Fatheree, Alexis B Carter, Lydia Contis, Bruce A Beckwith, Andrew J Evans, Avtar Lal, and Anil V Parwani. Validating whole slide imaging for diagnostic purposes in pathology: guideline from the college of american pathologists pathology and laboratory quality center. *Archives of Pathology and Laboratory Medicine*, 137(12):1710–1722, 2013.

[170] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[171] Srinivasa R Prasad, VR Narra, R Shah, PA Humphrey, J Jagirdar, JR Catena, NC Dalrymple, and CL Siegel. Segmental disorders of the nephron: histopathological and imaging perspective. *The British journal of radiology*, 80(956):593–602, 2007.

[172] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12556–12565, 2020.

[173] Yeping Lina Qiu, Hong Zheng, Arnout Devos, Heather Selby, and Olivier Gevaert. A meta-learning approach for genomic survival analysis. *Nature communications*, 11(1):1–11, 2020.

[174] Gwenolé Quellec, Mathieu Lamard, Pierre-Henri Conze, Pascale Massin, and Béatrice Cochener. Automatic detection of rare pathologies in fundus photographs using few-shot learning. *Medical image analysis*, 61:101660, 2020.

[175] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

[176] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[177] Prashanth Rawla. Epidemiology of prostate cancer. *World journal of oncology*, 10(2):63, 2019.

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

[179] Victor E Reuter. The pathology of renal epithelial neoplasms. In *Seminars in oncology*, volume 33, pages 534–543. Elsevier, 2006.

[180] Abtin Riasatian, Morteza Babaie, Danial Maleki, Shivam Kalra, Mojtaba Valipour, Sobhan Hemati, Manit Zaveri, Amir Safarpoor, Sobhan Shafiei, Mehdi Afshari, et al. Fine-tuning and training of densenet for histopathology image representation using tcga diagnostic slides. *Medical Image Analysis*, 70:102032, 2021.

[181] Stephanie Robertson, Hossein Azizpour, Kevin Smith, and Johan Hartman. Digital image analysis in breast pathology—from image processing techniques to artificial intelligence. *Translational Research*, 194:19–35, 2018.

[182] Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. Can contrastive learning avoid shortcut solutions? *Advances in neural information processing systems*, 34:4974–4986, 2021.

[183] Pau Rodríguez, Jordi Gonzalez, Guillem Cucurull, Josep M Gonfaus, and Xavier Roca. Regularizing cnns with locally constrained decorrelations. *arXiv preprint arXiv:1611.01967*, 2016.

[184] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[185] Sebastian Ruder. Transfer Learning - Machine Learning's Next Frontier. http://ruder.io/transfer-learning/, 2017.

[186] Arnout C Ruifrok, Dennis A Johnston, et al. Quantification of histochemical staining by color deconvolution. *Analytical and quantitative cytology and histology*, 23(4):291–299, 2001.

[187] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[188] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018.

[189] Saori Sakaue, Jun Hirata, Masahiro Kanai, Ken Suzuki, Masato Akiyama, Chun Lai Too, Thurayya Arayssi, Mohammed Hammoudeh, Samar Al Emadi, Basel K Masri, et al. Dimensionality reduction reveals fine-scale structure in the japanese population

with consequences for polygenic risk prediction. *Nature communications*, 11(1):1–11, 2020.

[190] Soad Samir, Eid Emary, Khaled El-Sayed, and Hoda Onsi. Optimization of a pre-trained alexnet model for detecting and localizing image forgeries. *Information*, 11(5):275, 2020.

[191] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016.

[192] Lukas Schott, Julius Von Kügelgen, Frederik Träuble, Peter Gehler, Chris Russell, Matthias Bethge, Bernhard Schölkopf, Francesco Locatello, and Wieland Brendel. Visual representation learning does not generalize strongly within the same domain. *arXiv preprint arXiv:2107.08221*, 2021.

[193] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[194] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. *Advances in neural information processing systems*, 28, 2015.

[195] Tiffany L Sellaro, Robert Filkins, Chelsea Hoffman, Jeffrey L Fine, Jon Ho, Anil V Parwani, Liron Pantanowitz, and Michael Montalto. Relationship between magnification and resolution in digital pathology systems. *Journal of pathology informatics*, 4, 2013.

[196] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[197] Seonguk Seo, Yumin Suh, Dongwan Kim, Geeho Kim, Jongwoo Han, and Bohyung Han. Learning to optimize domain specific normalization for domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 68–83. Springer, 2020.

[198] Andrea Sgarro. Informational divergence and the dissimilarity of probability distributions. *Calcolo*, 18(3):293–302, 1981.

[199] Fereshteh Shakeri, Malik Boudiaf, Sina Mohammadi, Ivaxi Sheth, Mohammad Havaei, Ismail Ben Ayed, and Samira Ebrahimi Kahou. Fhist: a benchmark for few-shot classification of histological images. *arXiv preprint arXiv:2206.00092*, 2022.

[200] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*, 2018.

[201] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014.

[202] Zhiqiang Shen, Zhuang Liu, Jianguo Li, Yu-Gang Jiang, Yurong Chen, and Xiangyang Xue. Object detection from scratch with deep supervision. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):398–412, 2019.

[203] Efrat Shimron, Jonathan I Tamir, Ke Wang, and Michael Lustig. Implicit data crimes: Machine learning bias arising from misuse of public data. *Proceedings of the National Academy of Sciences*, 119(13):e2117203119, 2022.

[204] Milad Sikaroudi, Benyamin Ghojogh, Fakhri Karray, Mark Crowley, and Hamid R Tizhoosh. Batch-incremental triplet sampling for training triplet networks using bayesian updating theorem. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7080–7086. IEEE, 2021.

[205] Milad Sikaroudi, Benyamin Ghojogh, Fakhri Karray, Mark Crowley, and Hamid R Tizhoosh. Magnification generalization for histopathology image embedding. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1864–1868. IEEE, 2021.

[206] Milad Sikaroudi, Benyamin Ghojogh, Amir Safarpoor, Fakhri Karray, Mark Crowley, and Hamid R Tizhoosh. Offline versus online triplet mining based on extreme distances of histopathology patches. In *International Symposium on Visual Computing*, pages 333–345. Springer, 2020.

[207] Milad Sikaroudi, Maryam Hosseini, Ricardo Gonzalez, Shahryar Rahnamayan, and HR Tizhoosh. Generalization of vision pre-trained models for histopathology. *Scientific reports*, 13(1):6065, 2023.

[208] Milad Sikaroudi, Shahryar Rahnamayan, and Hamid R Tizhoosh. Hospital-agnostic image representation learning in digital pathology. *arXiv preprint arXiv:2204.02404*, 2022.

[209] Milad Sikaroudi, Shahryar Rahnamayan, and HR Tizhoosh. Alfa–leveraging all levels of feature abstraction for enhancing the generalization of histopathology image classification across unseen hospitals. *arXiv preprint arXiv:2308.03936*, 2023.

[210] Milad Sikaroudi, Amir Safarpoor, Benyamin Ghojogh, Sobhan Shafiei, Mark Crowley, and Hamid R Tizhoosh. Supervision and source domain impact on representation learning: A histopathology case study. *arXiv preprint arXiv:2005.08629*, 2020.

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

[212] Robert A Smith, Vilma Cokkinides, and Harmon J Eyre. American cancer society guidelines for the early detection of cancer, 2006. *CA: a cancer journal for clinicians*, 56(1):11–25, 2006.

[213] Fabio A Spanhol, Luiz S Oliveira, Paulo R Cavalin, Caroline Petitjean, and Laurent Heutte. Deep features for breast cancer histopathological image classification. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1868–1873. IEEE, 2017.

[214] Fabio A Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7):1455–1462, 2015.

[215] Fabio Alexandre Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. Breast cancer histopathological image classification using convolutional neural networks. In *International Joint Conference on Neural Networks*, pages 2560–2567. IEEE, 2016.

[216] K Stacke, G Eilertsen, J Unger, and C Lundström. A closer look at domain shift for deep learning in histopathology. arxiv. *arXiv preprint arXiv:1909.11575*, 10, 2019.

[217] Karin Stacke, Gabriel Eilertsen, Jonas Unger, and Claes Lundström. Measuring domain shift for deep learning in histopathology. *IEEE journal of biomedical and health informatics*, 25(2):325–336, 2020.

[218] Juan Luis Suárez, Salvador García, and Francisco Herrera. A tutorial on distance metric learning: Mathematical foundations, algorithms and software. *arXiv preprint arXiv:1812.05944*, 2018.

[219] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[220] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.

[221] Swapna. Convolutional neural network: Deep learning. https://developersbreach.com/convolution-neural-network-deep-learning/, Jul 2021.

[222] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[223] Sairam Tabibu, P. K. Vinod, and C. V. Jawahar. Pan-Renal Cell Carcinoma classification and survival prediction from histopathology images using deep learning. *Scientific Reports*, 9(1), dec 2019.

[224] Sairam Tabibu, PK Vinod, and CV Jawahar. Pan-renal cell carcinoma classification and survival prediction from histopathology images using deep learning. *Scientific reports*, 9(1):1–9, 2019.

[225] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

[226] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

[227] Eu Wern Teh and Graham W Taylor. Learning with less data via weakly labeled patch classification in digital pathology. *arXiv preprint arXiv:1911.12425*, 2019.

[228] David Tellez, Maschenka Balkenhol, Irene Otte-Höller, Rob van de Loo, Rob Vogels, Peter Bult, Carla Wauters, Willem Vreuls, Suzanne Mol, Nico Karssemeijer, et al. Whole-slide mitosis detection in h&e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks. *IEEE transactions on medical imaging*, 37(9):2126–2136, 2018.

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

[230] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[231] Hamid Reza Tizhoosh and Liron Pantanowitz. Artificial intelligence and digital pathology: challenges and opportunities. *Journal of pathology informatics*, 9, 2018.

[232] Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. A deeper look at dataset bias. In *Domain adaptation in computer vision applications*, pages 37–55. Springer, 2017.

[233] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[234] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

[235] Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *Advances in Neural Information Processing Systems*, 33, 2020.

[236] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[237] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[238] Joaquin Vanschoren. Meta-learning: A survey. *arXiv preprint arXiv:1810.03548*, 2018.

[239] Mitko Veta, Paul J van Diest, and Josien PW Pluim. Cutting out the middleman: measuring nuclear area in histopathology slides without segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 632–639. Springer, 2016.

[240] Mitko Veta, Josien PW Pluim, Paul J Van Diest, and Max A Viergever. Breast cancer histopathology image analysis: A review. *IEEE transactions on biomedical engineering*, 61(5):1400–1411, 2014.

[241] Nicholas J Vogelzang and Walter M Stadler. Kidney cancer. *The Lancet*, 352(9141):1691–1696, 1998.

[242] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31, 2018.

[243] Bailin Wang, Mirella Lapata, and Ivan Titov. Meta-learning for domain generalization in semantic parsing. *arXiv preprint arXiv:2010.11988*, 2020.

[244] Fei Wang and Jimeng Sun. Survey on distance metric learning and dimensionality reduction in data mining. *Data mining and knowledge discovery*, 29(2):534–564, 2015.

[245] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[246] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[247] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2015.

[248] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

[249] Quan Wen, Jiazi Yan, Boling Liu, Daying Meng, and Siyi Li. A meta-learning method for histopathology image classification based on lstm-model. In *Tenth International Conference on Graphics and Image Processing (ICGIP 2018)*, volume 11069. International Society for Optics and Photonics, 2019.

[250] Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre Alvise-Rebuffi, Ira Ktena, Taylan Cemgil, et al. A fine-grained analysis on distribution shift. *arXiv preprint arXiv:2110.11328*, 2021.

[251] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[252] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[253] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[254] Yan Xu, Zhipeng Jia, Yuqing Ai, Fang Zhang, Maode Lai, I Eric, and Chao Chang. Deep convolutional activation features for large scale brain tumor histopathology image classification and segmentation. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 947–951. IEEE, 2015.

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

[256] Zheng Xu, Wen Li, Li Niu, and Dong Xu. Exploiting low-rank structure from latent domains for domain generalization. In *European Conference on Computer Vision*, pages 628–643. Springer, 2014.

[257] Yukako Yagi. Color standardization and optimization in whole slide imaging. In *Diagnostic pathology*, volume 6, page S15. Springer, 2011.

[258] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019.

[259] Jiawei Yang, Hanbo Chen, Jiangpeng Yan, Xiaoyu Chen, and Jianhua Yao. Towards better understanding and better generalization of few-shot classification in histology images with contrastive learning. *arXiv preprint arXiv:2202.09059*, 2022.

[260] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Universal domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2720–2729, 2019.

[261] Yaodong Yu, Heinrich Jiang, Dara Bahri, Hossein Mobahi, Seungyeon Kim, Ankit Singh Rawat, Andreas Veit, and Yi Ma. An empirical study of pre-trained vision models on out-of-distribution generalization. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

[262] Manit Zaveri, Shivam Kalra, Morteza Babaie, Sultaan Shah, Savvas Damskinos, Hany Kashani, and Hamid R Tizhoosh. Recognizing magnification levels in microscopic snapshots. *arXiv preprint arXiv:2005.03748*, 2020.

[263] Jing Zhang, Zewei Ding, Wanqing Li, and Philip Ogunbona. Importance weighted adversarial nets for partial domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8156–8164, 2018.

[264] Yijun Zhao, Marcello Chang, Robin Wang, Ianto Lin Xi, Ken Chang, Raymond Y Huang, Martin Vallières, Peiman Habibollahi, Mandeep S Dagli, Matthew Palmer, et al. Deep learning based on mri for differentiation of low-and high-grade in low-stage renal cell carcinoma. *Journal of Magnetic Resonance Imaging*, 52(5):1542–1549, 2020.

[265] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30(11):3212–3232, 2019.

[266] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[267] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[268] Ariana Znaor, Joannie Lortet-Tieulent, Mathieu Laversanne, Ahmedin Jemal, and Freddie Bray. International variations and trends in renal cell carcinoma incidence and mortality. *European urology*, 67(3):519–530, 2015.

Glossary

ALFA is the proposed approach in this thesis that leverages all levels of feature abstractions for generalization to an unseen domain/hospital. xv, xvi, 73, 74, 78–85, 88, 89

CAMELYON is a grand challenge in pathology organized by the Diagnostic Image Analysis Group (DIAG) and Department of Pathology of the Radboud University Medical Center (Radboudumc) in Nijmegen, The Netherlands. xiv, 44, 45, 56

DeCAF A Deep Convolutional Activation Feature for Generic Visual Recognition is the title of a paper that introduced the idea of transfer learning for the first time. 15

DenseNet121 DenseNet121 is a deep learning model known for its unique 'dense connection' structure where each layer is connected to every other layer within each dense block. This 121-layer architecture improves information and gradient flow, mitigates vanishing gradient issues, and enhances computational efficiency. It's particularly effective for image classification tasks due to its robust feature extraction capabilities. 43

Grad-CAM Gradient-weighted Class Activation Mapping is a technique that produces a heat map overlay highlighting important regions in an image for a deep neural network's prediction by leveraging gradients of the target class score with respect to the feature maps of the final convolutional layer. It helps visualize and understand the decision-making process of CNNs in computer vision tasks. xiv, 52, 54, 56, 57

HA Hospital-Agnostic Image Representation Learning approach proposed in this thesis and already been presented at IEEE EMBC 2022 conference. xv, 68–70, 78, 80–82, 84, 85

histomorphologic The use of histology to study the morphology of cells. 2

HMD is a combined repository, bringing together data from both Harvard and MD Anderson Cancer Centers. It is used for the Renal Cell Carcinoma (RCC) subtyping task within the scope of this thesis, providing a benchmark to assess and compare various methodologies. xv, 66, 69, 70, 79, 80, 84

ImageNet is a large-scale, diverse dataset used primarily for object recognition tasks in the field of machine learning and computer vision. It consists of over 14 million images labeled across 20,000 categories, providing a wide-ranging resource for training and benchmarking deep learning models.. 11, 15, 26, 27, 43, 47, 56, 58, 59, 79

KimiaNet is a deep learning network that is uniquely tailored for histopathology image representation. It's built on the DenseNet architecture and includes four dense blocks. The network has been refined and trained using a variety of configurations of histopathology images, displaying superior results as a feature extractor for these images compared to both the original DenseNet and smaller Convolutional Batch Normalized ReLU (CBR) networks. xiv, xvii, 27, 43, 47, 49–52, 54–59

latent space representation Any encoder like CNN or transformers aims to encode inputs across multiple layers prior to use for a downstream task (classification, regression, image reconstruction). In other words, any deep model acts like a mapping function that projects the input onto latent space through some computational layers. The encoded version of the deep models' input is called the latent space representation which has a lower dimension with respect to the input space. xiii, 13–15, 22, 31, 32, 35–37, 71

PACS is a popular benchmark for studying domain generalization. It consists of images from four distinct domains: Photo, Art painting, Cartoon, and Sketch (hence the acronym PACS). Each domain contains images spanning seven common classes. This dataset provides a diverse and challenging environment for evaluating algorithms' capacity to learn domain-invariant features and generalize well across varying domains. xv–xvii, 73, 79–83

ResNet18 is a CNN model, a part of the ResNet series (Residual Networks) proposed by Microsoft Research. It consists of 18 layers, including convolutional layers, pooling layers, and fully connected layers, and utilizes shortcut connections or "skip connections" to address the problem of vanishing gradients in deep neural networks, enabling efficient training of much deeper networks. 43, 47, 79