Building Computer-aided Diagnostic Models for Biopsy Images Using Minimally Curated Datasets

by

Joseph Vincent Valdez Pulido

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Systems and Information Engineering

of the

University of Virginia, Charlottesville, Virginia

School of Engineering

Committee:

Professor Michael Porter, PhD, Chair
Professor Donald E. Brown, PhD, Co-adviser
Professor Laura Barnes, PhD, Co-adviser
Doctor Sana Syed, MD
Professor Peter Beling, PhD

Spring 2021

The dissertation of Joseph Vincent Valdez Pulido, titled Building Computer-aided Diagnostic Models for Biopsy Images Using Minimally Curated Datasets, is approved:

Chair _____ Date _____

_____ Date _____

_____ Date _____

_____ Date _____

_____ Date _____

University of Virginia, Charlottesville

Building Computer-aided Diagnostic Models for Biopsy Images Using Minimally Curated Datasets

Abstract

Building Computer-aided Diagnostic Models for Biopsy Images Using Minimally Curated Datasets

by

Joseph Vincent Valdez Pulido

Doctor of Philosophy in Systems and Information Engineering

University of Virginia, Charlottesville

Convolutional neural networks (CNNs) perform well on many biopsy datasets and show promising signs of becoming part of the process to augment physicians' workflow to diagnose diseases. Training these medical models typically involves the tedious task of curating large-scale medical datasets which involves activities like collecting, storing, and annotating data. In general, these data curation tasks are usually the most time consuming and resource intensive portion of the model development processes. This cost is exacerbated in the medical field where 1) annotation labor costs are high, and 2) disease samples are scarce. This body of work aims to reduce data curation costs by examining and developing methods that aim to alleviate the problems of building CNNs trained on partially annotated or imbalanced datasets. First, I will examine the performance of *semi-supervised* methods that increase prediction performance by leveraging a large unlabeled dataset, along with a smaller labeled dataset. I assess the impact of having noisy samples in the unlabeled data which is common in biopsy tissue data. To decrease the effects of this noise, I examine the effects of applying semi-supervised co-teaching methods. Next, I analyze the performance of *class imbalance* methods on the task of grading cancerous biopsies. I show that state-of-the-art class imbalance methods perform sub-optimally due to rare 'polarized features' inherent in many biopsy cancer grading tasks–where cancer patterns manifest only at the tail-end of the cancer progression. By improving these two areas of research, this work aims to decrease the cost of curating biopsy datasets and promote the use of CNNs on many medical tasks under resource constrained settings.

To Sarah.

Contents

List of Figures

List of Tables

Acknowledgments

Throughout the writing of this dissertation, I have received a great amount of support and assistance:

I would first like to thank my supervisors, Professor Donald E. Brown and Doctor Sana Syed, whose expertise was invaluable in formulating the research questions and methodology. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level. I would also like to thank their patient support and for all the opportunities I was given to further my research.

I would also like to thank the Johns Hopkins University, Applied Physics Lab who funded and supported my doctoral studies. I especially want to thank Dean Fisher for his support and belief in my abilities to complete this dissertation.

I would like to acknowledge my colleagues from the Gut Intelligence Lab for their wonderful collaboration. I would particularly like to acknowledge William Adorno and Yash Sharma for their invaluable feedback and reviews to polish my work.

In addition, I would like to thank my parents for their ongoing inspiration and counsel. You are always there for me. Finally, I could not have completed this dissertation without the support of my wife, Sarah, who provided stimulating discussions, happy distractions, and a sympathetic ear to rest my mind outside of my research. I couldn't have done this without your support.

Chapter 1

Introduction

1.1 Overview

This chapter provides the introduction to this dissertation. Section 1.2 describes the motivation for the need to decrease the cost of curating medical biopsy datasets. Section 1.3 describes the problem formulation and our proposed approach to decrease data curation costs. Sections 1.4 and 1.5 describe the prerequisite background for the main body of the dissertation: Convolutional Neural Networks and biopsies. Section 1.7 describes the purpose, scope, and the list of scientific contributions of this dissertation. Lastly, Section 1.8 describes the organization of the main body of this dissertation.

1.2 Motivation

Deep Convolutional Neural Networks (CNN) are potential tools to assist humans in tasks requiring laborious and monotonous visual inspections due to their predictive performance in many computer vision tasks. For example, in clinical diagnosis involving the examination of histopathological biopsies, the current standard of care involves highly-trained pathologists visually detecting and grading lesions of tissue samples. CNNs are expected to be used as first-pass filtering assistants to highlight areas of potential disease. Many medical researchers see these models as tools to alleviate pathologists' workload that allow them to attend to more cases.

The computing resources to train and store large-scale CNNs have become increasingly cheaper and more democratized as more industries have lower barriers to build their own custom deep neural networks. Despite the wide availability and improvement of enabling technologies like specialized computer hardware (e.g. GPUs and cloud computing), the largest cost has still come from activities relating to the curation (i.e. collection, storage, and annotation) of well-manicured datasets for training and evaluating these models. This cost is exacerbated in the field of medical imaging where 1) labor cost for annotation is high, and 2) examples of diseases disproportionately are the minority of samples collected. These

curation costs contribute to the high obstacle to apply fully-supervised machine learning on medical datasets. Lowering the cost of data curation activities may lead to the proliferation of deep learning technologies in medical imaging, resulting in both lightening the burden of physicians and deceasing the time to diagnosis.

1.3 Problem Formulation

This body of work contributes to the development and evaluation of computer-aided diagnostic models that bypass the need to 1) expensively annotate all available medical imaging data, and 2) collect more data of rare, abnormal cases. Thus, this body of work addresses two challenges common in machine learning applied to medical imaging to palliate the heavy cost of data curation:

- The first major challenge is that the necessary annotation labor required to produce high-quality diagnostic models are usually more complicated, expensive, and time consuming compared to general knowledge datasets (e.g. person identification, pose estimation, etc.). Recruiting highly trained pathologists who are able to interpret these slides is expensive due to the time investment necessary to annotate data. This laborious process is costly as they must painstakingly label every sample of a large dataset, which takes valuable time away from their practice. This body of work proposes the use of **semi-supervised learning methods** (SSL) that leverage the use of learning from both a large unlabeled dataset and a smaller labeled dataset. Using SSL methods, one could avoid the need to annotate all available data. For example, using this paradigm, a model could be trained with only using a small fraction (i.e. 10-25%) of the dataset.
- The second major challenge is that medical data of rare diseases are difficult to acquire. In medical datasets, data are predominately composed of "normal" samples with only a small percentage of "abnormal" samples, leading to class imbalance problems. CNN's are susceptible to these imbalance problems as they may only learn patterns from the majority classes. This body of work proposes the application of **class imbalance methods** designed to train CNNs to avoid the biased learning of the patterns in the majority classes. Using methods addressing class imbalance, one does not need to search for more rare samples for the sake of balancing a dataset to improve the performance of the model. This ultimately lowers the cost of building these CNNs in the biopsy setting.

In the following sections, I will outline the prerequisite background for the topics in this dissertation. First, in Section 1.4, I will briefly introduce CNNs. Next, in Section 1.5 and 1.6 , I will describe biopsy images and their unique properties that make them different from general images commonly studied when analyzing CNNs performance.

1.4 Convolutional Neural Networks

CNNs are currently the leading solution for computer vision tasks as they have an ability to achieve almost human-like performance on a wide variety of tasks like object recognition [79, 30, 71, 32], image captioning [89], and pose estimation [87]. Using hierarchically organized convolutional operators, CNNs extract spatially preserved features from the input image. For an object recognition task, CNNs extract hierarchical patterns in images to assemble complex visual patterns (e.g. eyes, wheels, etc.) composed of smaller and simpler patterns (e.g. texture, edges, etc.). The convolutional parameters of a CNN are trained through an iterative trial-and-error process called back propagation [72]. At a very high-level, the back-propagation algorithm, first, calculates the gradients of the error with respect to weights in the network and, then, uses gradient descent to update parameter values to optimize a pre-determined objective function. Figure 1.1 shows how each layer transforms their respective input into a more abstracted representation which allows the network to ultimately learn important features to accomplish a task. These models achieve state-of-the-art results on large-scale benchmark datasets like ImageNet [47] for the classification of 1000 objects, and MS-COCO [55] for object detection and segmentation of 80 objects. Although these models perform well on these benchmark computer vision datasets, developing the large-scale datasets for training and evaluating CNNs involved the aggregated efforts of multiple participants to annotate about 1.5M and 200K images for ImageNet and MS COCO, respectively.

1.5 Biopsy Classification

CNNs have been successfully applied to medical imaging applications, from histology [40, 39] to MRI [61]. For this body of work, the focus will be on the analysis of biopsies– medical tests involving the extraction of sample tissues for visual examination to determine the presence or severity of a disease. Tissue is extracted by a surgeon and is sent to a laboratory to process the extracted samples by cutting a microscopically thin slice of the tissue mass and attaching the slice to a glass slide. The slide is then treated with dyes that stain the tissue, allowing the individual patterns (e.g. nucleus and cell boundaries) to be more discernible. The slide is visually examined by pathologists to find evidence of possible cancerous and inflammatory conditions. Figure 1.2 (left) depicts an example of an esophageal and prostate biopsy. Pathologists look for evidence of disease by searching for microscopic visual disease features apparent on the slide. For example, to monitor the progression of esophageal cancer, pathologists classify tissue into three categories: squamous (normal), barrett's non-dysplastic, and dysplastic, by order of severity. As another example, to inform physicians on how to treat prostate cancer, pathologist grade cancerous tissue into 3 Gleason scores [35]. Figure 1.2 (right) shows examples of Gleason 3, 4, and 5, by order of severity.

In the past, biopsies were visually analyzed under a microscope; however, current trends point towards digitization of these images. Although not yet the global standard practice,

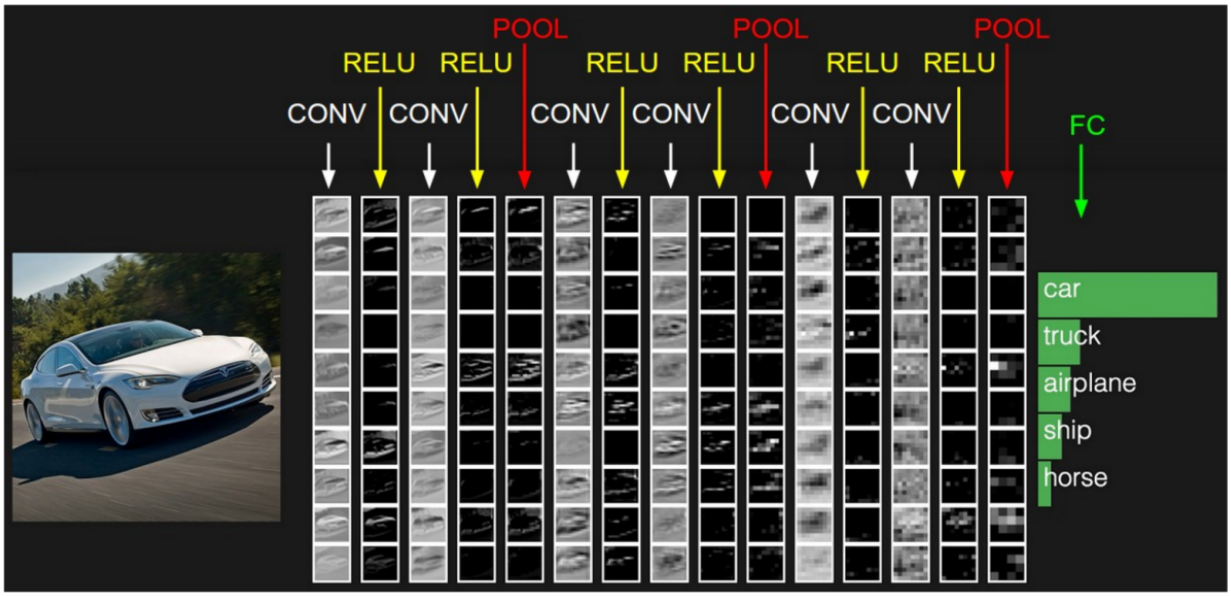

Figure 1.1: Visualization of convolutional architecture activations [1]. Beginning with the input image (left), each sequential layer of a CNN applies non-linear convolutional transformations to produce spatially organized features with higher abstractions in order to perform a classification (right).

the benefits of digitization include easy transfer and storage of data. Images of biopsies, called whole-slide images (WSI), are gigapixel-sized images that allow pathologists to zoom in (to examine small details of a tissue) and zoom out (to understand the context of certain microscopic patterns). Machine learning practitioners inadvertently benefited from such a transition as the digitization of these slides readily opened the door for the application of computer vision and image processing techniques, like CNNs, to various histological challenges [31, 21, 56, 45, 44, 62, 15, 92, 25, 5, 68].

The medical imaging community sees CNNs as a promising tool for automatic computer-aided diagnosis of biopsy images. They have been applied to detect numerous diseases such as breast cancer [58, 3], liver cancer [52], lung cancer, [90] and esophageal cancer [86]. All of these studies have been applied in research settings where CNNs are shown to achieve human-like performance.

1.6 Unique Characteristics of Biopsy Datasets

This section details some of the unique characteristics typically associated with biopsy imaging relative to generic computer vision datasets (e.g. ImageNet [47] and MS COCO [55]).

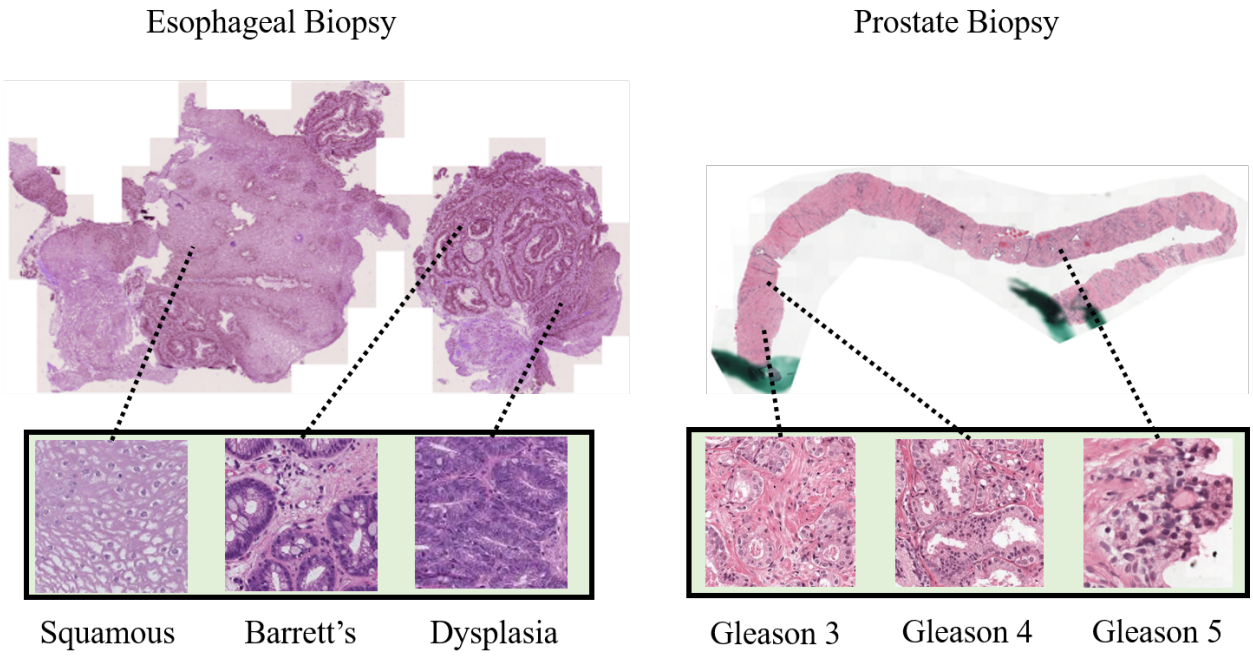

Figure 1.2: An example of esophageal (left) and prostate (right) biopsies that pathologists analyze to detect disease. Pathologists visually scan areas for patterns which indicate evidence of disease. These gigapixel whole-slide images contain visual information of microscopic detail of tissue samples extracted from a patient; however, areas of interest may only be found in a few thousand pixels. To analyze esophageal biopsies for the sake of monitoring the esophageal cancer, pathologists examine ares of cancer precursors [77]. To analyze prostate cancer, pathologists grade areas of cancer using the Gleason grading system [20]. The patches are extracted at 40x magnification level. The esophageal patches is extracted at 1000x1000 and the prostate patches are extracted at 512x512.

Gigapixel-sized images

Digitized biopsy slides are high resolution images that are much larger than standard images commonly used in computer vision applications. These high resolution sizes are prohibitive for the application of neural networks as 1) resizing these images would destroy microscopic patterns important to the diagnosis, and 2), if resizing was not performed, off-the-shelf GPUs do not carry enough memory to store the parameters of a large model required by a gigapixel-sized input.

To ameliorate the gigapixel problem, the common practice is to perform "patching" operations by subdividing the WSI into smaller images–cropped in a sliding window manner– to use them as input data to the CNN (See Figure 1.3). These patches should be small enough to fit into GPU memory and have enough visual descriptors to carry patterns present in the

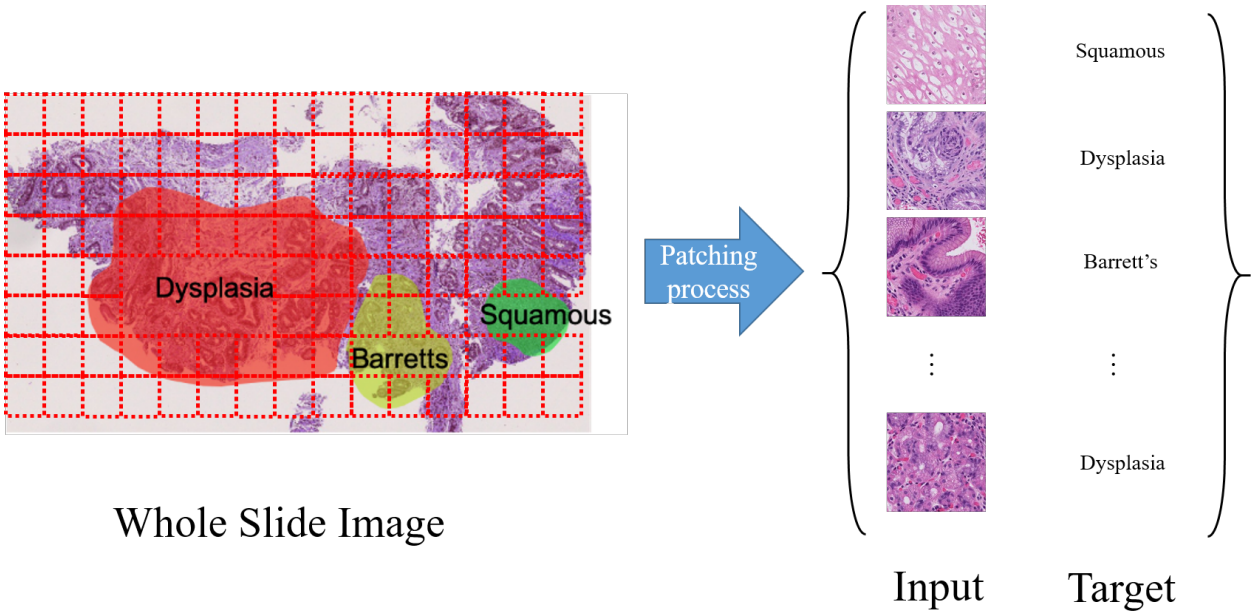

Figure 1.3: One of the common challenges in computational pathology is the processing the gigapixel size of a single whole-slide image. Using the highest magnification, a single image can contain several billion pixels, while the area of interest can be as small as a few thousand pixels. To apply a deep learning classifier, the whole-slide image has to be divided into several thousand tiles–a process called "patching." In this figure, a whole-slide image is annotated by a pathologist into segments. The slide is then subdivided into patches. Using the segmentation information, each patch is then labeled a certain class if the patch overlaps with a segmentation class. The subdivided patches and targets are used as training data for convolutional neural networks.

diseases. This method has shown to perform well on classification and detection tasks of biopsy slides [33, 90, 86, 95, 16, 13].

One benefit to the patching approach is the generation of multiple images, which is also conducive to the training of data-hungry CNNs. One WSI could produce hundreds, or even thousands, of training images.

Open-set, multi-class noisy images

Biopsies contain two types of visual noise: 1) open-set noise and 2) multi-class noise. Open-set noise are areas where the biopsy contains tissue structures that are not relevant to the context of the problem. For example, in biopsies for the examination of esophageal cancer in Figure 1.4, a biopsy could inadvertently sample gastric (stomach) tissue or muscular tissue. This is called open-set noise because these portions of the biopsy are *outside* the set of

classes-in-question. In the biopsy domain, if there is a high occurrence of open-set patches contained in the training data, CNNs will inadvertently overfit to these open-set images [93]. Thus, it is important to correctly handle these patches during training. The second source of noise is called multi-class noise which originates from a biopsy sample containing multiple classes present in the image. Because diagnoses usually describe only the most severe class, information of the occurrence of other classes disappear at the diagnosis (i.e. physicians usually only look for the most severe label). When patching these biopsies, one must be aware to annotate these patches carefully into its proper class. Figure 1.5 shows an example of multi-class biopsy slide with of an esophageal tissue.

Imbalanced datasets

More often than not, real-world datasets have classes that are severely underrepresented in sample size relative to others. This is especially true with general medical datasets where some diseases are rare and collecting more data is difficult. Imbalanced datasets could be detrimental to classifiers like CNNs. When trained on a highly imbalanced dataset, a classifier has a tendency to pick up the patterns in the most popular classes and ignore the least popular ones.

1.7 Purpose and Scope

This body of work aims to alleviate the data curation cost of biopsy datasets by addressing the annotation costs and class imbalances caused by rare diseases common in biopsy applications. To this end, this dissertation contains the following four contributions:

- Contribution 1: I analyze the performance of modern computer vision SSL techniques on esophageal [70, 26] and prostate biopsies.
- Contribution 2: I compare and contrast the performance of leading SSL methods on biopsy images. I highlight that current SSL research ignores the inherent characteristics of biopsy images–more specifically, the noise contained in biopsy image [70].
- Contribution 3: In order to improve SSL techniques on noisy biopsy images, I evaluated the application of co-teaching [27, 53] within a SSL setting on prostate biopsies. I introduce CoMixMatch which extends the MixMatch [8] semi-supervised learning technique.
- Contribution 4: I analyzed the effects of two-stage methods [38] on prostate biopsy images for the task of cancer grading. I discover a property unique to biopsy images which I call "polarized features" which organizes cancer features exclusively at the tail-end (i.e. minority classes) of the cancer progression. I show that these polarized features cause sub-optimal performance when applying state-of-the-art two-stage methods [38] on imbalanced datasets.

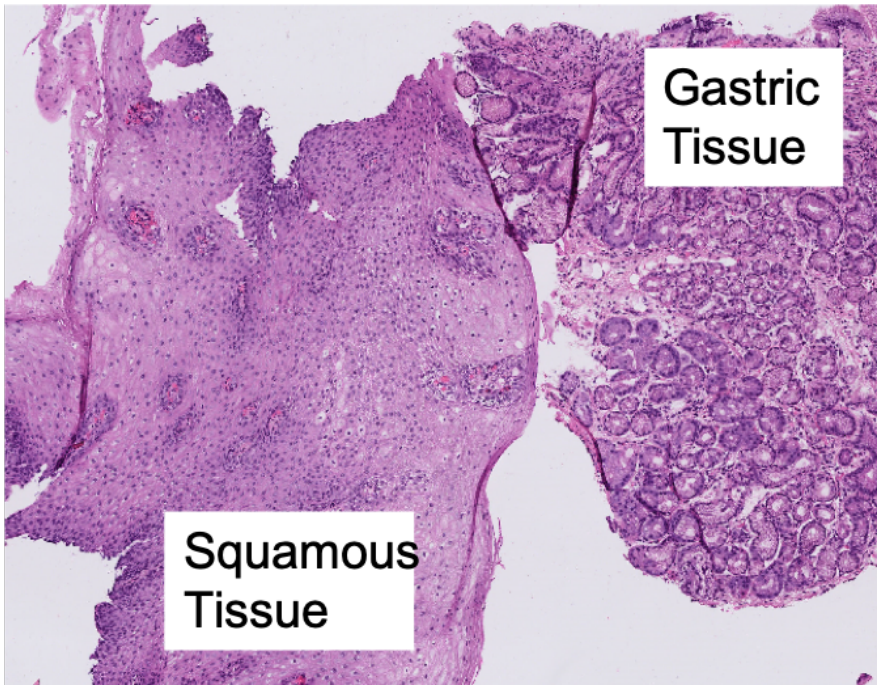

Figure 1.4: An example of open-set noise present in an esophageal biopsy that included gastric tissue which is irrelevant to the esophageal cancer analysis

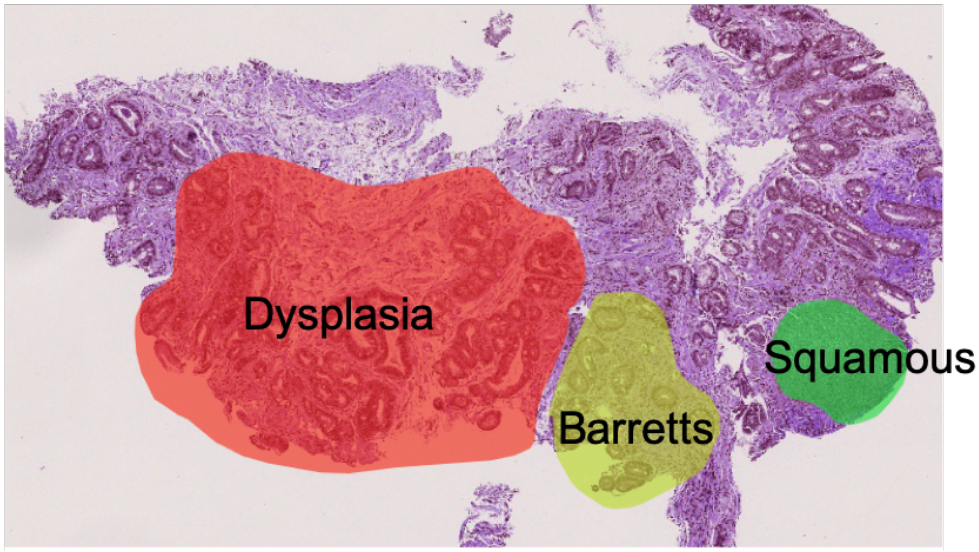

Figure 1.5: An example of multi-class noisy present in a biopsy containing three classes of tissue.

1.8 Organization of Dissertation

This dissertation is organized as follows: Chapter 2 lists the pertinent literature related to this work where I detail previous work on deep learning techniques on biopsy classification, modern SSL methods, and class imbalance methods. In Chapter 3, I compare and contrast two leading SSL methods, MixMatch [8] and Fixmatch [81], on a esophageal biopsy dataset (Contribution 1). I show that MixMatch is more robust to the noise inherent in biopsy images (Contribution 2). In Chapter 4, I study SSL techniques on the prostate dataset (Contribution 1). I study the impacts of using co-teaching methods, a multi-model method that aims to decrease the impacts of noisy biopsy datasets (Contribution 3). In Chapter 5, I study the impacts of biopsy datasets on two-stage imbalanced methods that aim to decrease the impacts of rare disease samples. In this chapter, I empirically show that current two-stage methods perform sub-optimally due to polarized features inherent in cancer grading tasks of biopsy images (Contribution 4). Finally, I conclude with Chapter 6 to summarize this body of work, address limitations, and point to possible future directions.

Chapter 2

Literature Review

2.1 Overview

This chapter outlines the pertinent work in biopsy classification (Section 2.2), semi-supervised learning (Section 2.3), class imbalances (Section 2.4) and training with noisy data (Section 2.5).

2.2 Whole-slide Image Classification

In conventional clinical diagnosis, histopathological examination of biopsy samples is visually analyzed under light microscopy. Whole slide images (WSIs) are the digitized counterparts of glass slides obtained via specialised scanning devices, and they are considered to be comparable to microscopy for primary diagnosis[66]. The growing use of digitized WSIs to examine biopsies opened the door to apply computer vision and image processing techniques. In particular, deep convolutional neural networks (CNNs) have shown state-of-the-art results in a large number of computer vision applications like whole-slide histopathology for lung cancer [25, 92, 15, 62], prostate cancer [56], colon cancer [44], breast cancer [5] and skin cancer [21]. These results highlight the potential large benefits that could be obtained when deploying deep learning-based tools to aid pathologists' diagnosis workflow systems especially for increasing first-pass screening efficiency.

2.3 Semi-supervised Learning

There are a wide range of SSL techniques [11, 88]: Researchers have explored transductive models [22, 37, 36], graph-based methods [97, 7, 57], and generative models [6, 50, 73, 14]. For this work, we will concentrate this review on the pertinent papers that address *consistency regularization* and *entropy minimization* which have shown to be highly effective in the image domain [8, 9, 81].

Several works have explored the use of SSL on histology images. Lu et al. [60] used a two-stage approach using self-supervised contrastive learning and a multi-instance attention module to the task of binary classification of breast cancer histology images. Peikari et al. [69] used a "cluster-then-label" approach finding high density areas of unlabeled clusters then using these clusters to train a support vector machine (SVM) to learn a decision boundary through low density areas. This approach used a bag-of-words descriptor to represent each patch. To the best of my knowledge, this body of work is the first to examine modern SSL methods using consistency regularization and entropy minimization.

Consistency regularization

Consistency regularization is a constraint that forces models to produce consistent inferences despite the application of multiple input transformations. In the SSL setting, an unlabeled example must adhere to a single class no matter how it is augmented.

The Π-model [49] adheres to a consistency regularization by forming a consensus prediction of unknown labels by using the model at different times of training (called temporal ensembling), and under different input augmentation conditions. This ensemble includes a committee of historical "snapshots" of the model from different times of the training phase. Every model in the ensemble predicts a sample and the model is penalized if there are inconsistencies between the predictions.

Mean Teacher simplified [84] uses two networks, one called the student, and the other is called the teacher. At each training step, the same sample is used for training both models. A consistency cost is calculated between the student and the teacher which penalizes the deviation between the two. The consistency cost is then added to the standard classification cost (i.e. cross-entropy). At each iteration, the optimizer adjusts the student model weight, while the teachers weight is the exponential moving average of the student weight.

Entropy minimization

Entropy minimization is a constraint that assumes points close to one another should belong in the same class. Bengio [7] formulated an optimization scheme that trains a classifier producing low-entropy predictions on unlabeled data by having a decision boundary that avoids passing through clusters.

Virtual Adversarial Training (VAT) [65] is a technique that computes a minimal adversarial noise (i.e. an additive perturbation) which affects the output of the class distribution. It has been used in improving supervised learning performance and unsupervised clustering. VAT uses a virtual adversarial training loss which penalizes a model that is sensitive to adversarial perturbations by generating training inputs such that the perturbation changes the classifier's predicted label. VAT uses no label information and the perturbation is applied using just the model outputs. The perturbation is generated such that output of the perturbed input is different from the model output of the original input (as opposed to the

ground truth label). Applying the VAT loss on the unlabeled set and the supervised loss on the labeled set gives a boost in testing accuracy.

Pseudo-label [51] is a simple technique which uses a learning procedure that lets a model first "guess" the class of the unlabeled dataset. The unlabeled samples along with their pseudo-labels are mixed in with the labeled training set and are treated as conventional training examples.

This work is mostly build upon MixMatch [8] and FixMatch [81] which are SSL techniques that unifies the two conditions of consistency regularization and entropy minimization. MixMatch uses consistency constraints by averaging inferences from multiple transformations of a single image and enforcing a "sharpening" function to accentuate the dominant class. This sharpened prediction then becomes the pseudo-label from which the model learns. On the other hand, FixMatch uses consistency constrains by enforcing the model to adhere to a single input under various transformations like RandAugment [17] and CTAugment [9].

2.4 Class Imbalance

Imbalanced datasets have been studied for decades. The over arching theme of the available solutions is to avoid the biased learning of the features within the majority classes. There are three main pathways to address class imbalances: rebalancing (Section 2.4), class-senstive losses (Section 2.4), and transfer learning (Section 2.4).

Data rebalancing

Researchers have proposed to re-sample the dataset to achieve a more balanced data distribution. These methods include over-sampling [12, 28, 29] for the minority classes (by adding copies of data or synthetic data), undersampling [19, 85] for the majority classes (by removing data), and class-balanced sampling [78, 63] based on the number of samples for each class.

Class-sensitive losses

Another approach involves assigning different losses to different training samples for each class. The loss can vary at class-level for matching a given data distribution and improving the generalization of minority classes [18, 42, 10, 41, 34]. These loss functions are usually weighted using an inverse class frequency. As the number of classes increase, the class weights decrease. A more fine-grained control of the loss, differentiated between easy and hard classes, can also be achieved at sample level using focal loss [54].

Transfer learning

In the deep learning era, there are exciting advances proposing solutions to the problem of imbalances that scale to datasets that mirror the expanse of the real world. These long-tailed datasets [59] have classes that follow a Zipfian distribution where a minority of the classes hold a majority of the samples and a majority of the classes hold a minority of the samples. To tackle the problem of long-tailed datasets, Kang et al. [38] proposed an effective method, called Decoupling, that follows a two-stage approach by detaching the process of learning features and learning discriminators. Transfer-learning based methods address the issue of imbalanced training data by transferring features learned from head classes with abundant training instances to under-represented tail classes. This successful paradigm for tacking imbalances have been extended to many other techniques [83, 96, 83, 57].

2.5 Noisy Data

Due to the inherent noise in biopsy images, this body of work closely relates to open-set recognition [24] and learning from noisy labels (or noisy supervision) [2].

In open-set recognition, models are trained on known classes and must infer if a class is outside the set of known classes. Approaches include SVM-methods like Scheirer et al. (2012) [75] who proposed 1-vs-Set support vector machine, which incorporates an open space risk term in modeling to account for the space beyond the reasonable support of known classes and Scheirer et al. (2014) [74] introduce an open set recognition model called compact abating probability (CAP), where the probability of class membership decreases in value as points move from known data toward open space. Work in open-set recognition train models on known samples and test on potentially open-set sample. In the SSL biopsy application, open-set data can also be in the training data.

Our work closely relates to Wang et al. [91] which learns a deep learning model despite the presence of significant open-set noise by first detecting noisy samples iteratively and using a contrastive loss to learn a metric that pushes noisy samples away from clean samples in a metric space. However, Wang et al. only learns from a fully-supervised setting. Co-teaching [27] uses a two-model method where each model dictates which of the samples to train. DivideMix [53] extends the use of co-teaching [27] to train a fully-supervised model against noisy labels (i.e. flipped labels). In this chapter, apply many DivideMix principles to address issues in open-set SSL.

Chapter 3

Semi-Supervised Classification of Esophageal Biopsies on Noisy, Gigapixel Histology Images

3.1 Overview

One of the greatest obstacles in the adoption of deep neural networks for new medical applications is that training these models typically require a large amount of manually labeled training samples. In this chapter, we investigate the semi-supervised scenario where one has access to large amounts of unlabeled data and only a few labeled samples. I study the performance of MixMatch and FixMatch–two popular semi-supervised learning methods–on a histology dataset. More specifically, I study these models' impact under a highly noisy and imbalanced setting. The findings here motivate the development of semi-supervised methods to ameliorate problems commonly encountered in medical data applications.

3.2 Background

Convolutional Neural Networks (CNN) have been the dominant framework in many computer vision tasks. The computing resources needed to train large scale CNN have become increasingly cheaper and more democratized as the barrier to train custom deep neural networks is lowered. Today, some of the larger costs have now come from activities relating to the annotation of datasets for training and evaluating these models. These costs are exacerbated in the field of medicine where experts' time is costly. This presents a high obstacle to apply fully-supervised machine learning techniques that requiring well-curated and fully annotated datasets.

In order to circumvent a fully-supervised model, researchers turn to techniques like semi-supervised learning (SSL) to minimize the annotation requirements to build comparable models. These learning techniques adapt to an environment where one has a small amount

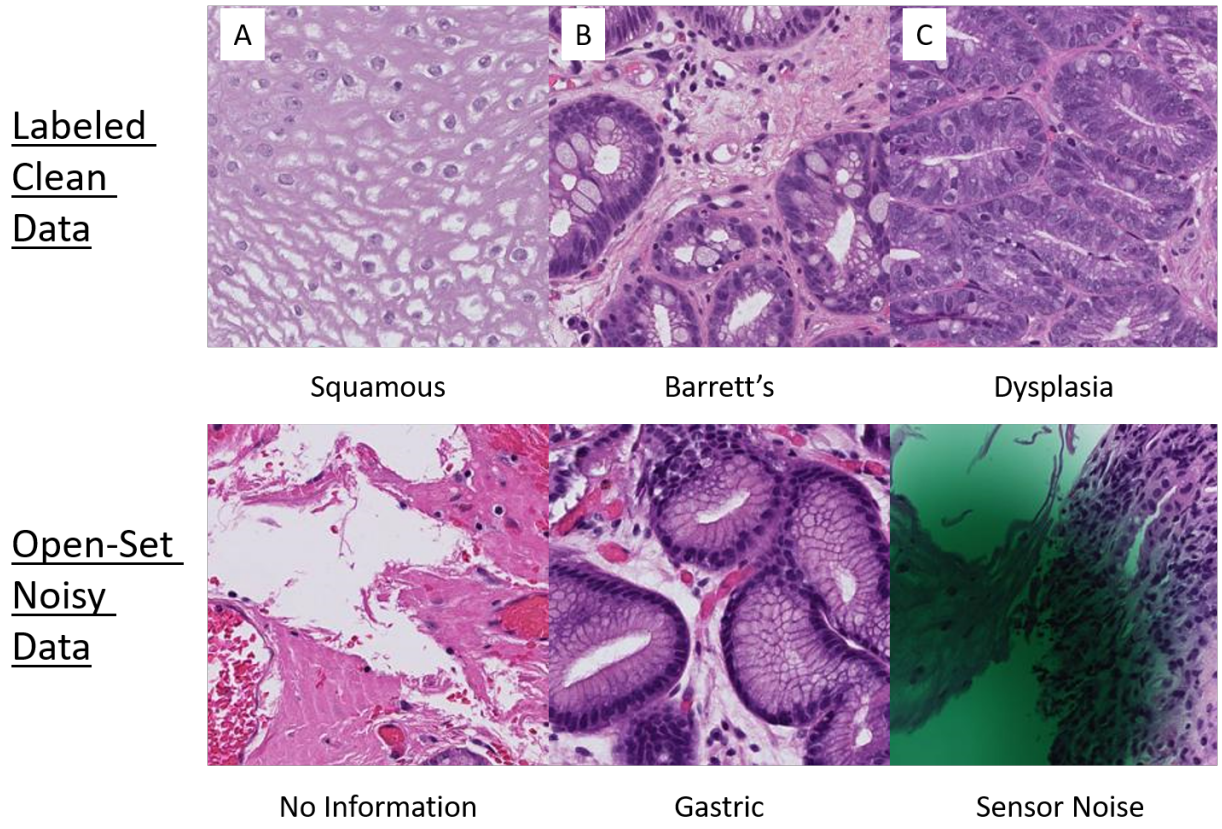

Figure 3.1: A) An example of normal squamous tissue of the esophagus, identified by flat, stratified cells. B) An example of non-dysplastic Barrett's esophagus, characterized by large white goblet cells filled with mucus and ovoid glands reminiscent of intestinal tissue. C) An example of dysplasia of the esophagus in which nuclei become more prominent with varying sizes and shapes (pleomorphism) and glands become more crowded. The bottom three examples are instances of open-set data which are data points that do not belong to any of the three classes in-question. They can include patches that add no information, tissue of a different type (e.g. gastric and muscular tissue), and areas of the image that contain sensor noise.

of labeled data and a larger proportion of unlabeled data. Recently, there has been a surge in state-of-the-art performance in SSL methods using MixMatch [8] and FixMatch [81]. Both techniques rely on pseudo-labeling (guessing unknown labels for training) and data augmentation to tackle semi-supervision; however, they diverge on the manner in which they perform these procedures.

Although these methods are empirically successful in general computer vision SSL tasks, they have not been examined under conditions common in the field of histology where class imbalances and noisy samples are common. In this study, I explore the performance of FixMatch and MixMatch on a semi-supervised histological setting. The contribution of this work is two fold: Firstly, I apply modern SSL methods on the task of detecting disease patterns by training a multi-class model using only a few labeled images while leveraging the use of a larger amount of unlabeled images. For our use case, I will be applying SSL methods on a histology dataset especially curated for the purpose of detecting esophageal cancer's precursors: dysplasia, Barrett's, and squamous tissue. Lastly, I will study the effects of imbalanced datasets on the two SSL methods.

3.3 Methods

This section will introduce MixMatch [8] and FixMatch [81]. Then I will analyze the main difference between the two models by their unique procedures of pseudo-labeling, data augmentation, and unlabeled loss.

MixMatch

MixMatch, a recently developed SSL technique, is designed around the common idea that models can organize the unlabeled samples using ideas of entropy minimization (which encourages the model to have consistent labels for similar images) and consistency regularization (which encourages the model to have consistent labels for a single sample despite small changes to the image). At a very high level, MixMatch first performs pseudo-labeling, where a model guesses the low-entropy class of all unlabeled samples; then, MixMatch creates synthetic data combining labeled samples and unlabeled samples using a technique called MixUp [94]. The iterative use of pseudo-labeling and MixUp allows the model to effectively learn from unlabeled data using optimization methods of traditional fully-supervised learning techniques like backpropagation. This section further details the pseudo-labeling and MixUp procedures of MixMatch.

Pseudo-labeling and Loss Function

Let $\mathcal{X} = (x_i, y_i), i \in \{1, \ldots, N\}$ be the set of N labeled samples where x_i and y_i are the input and label of the i^{th} sample respectively. And let $\mathcal{U} = u_j, j \in \{1, \ldots, M\}$ be the set of M unlabeled samples where u_j is the j^{th} sample. For an input x_i or u_j, MixMatch performs

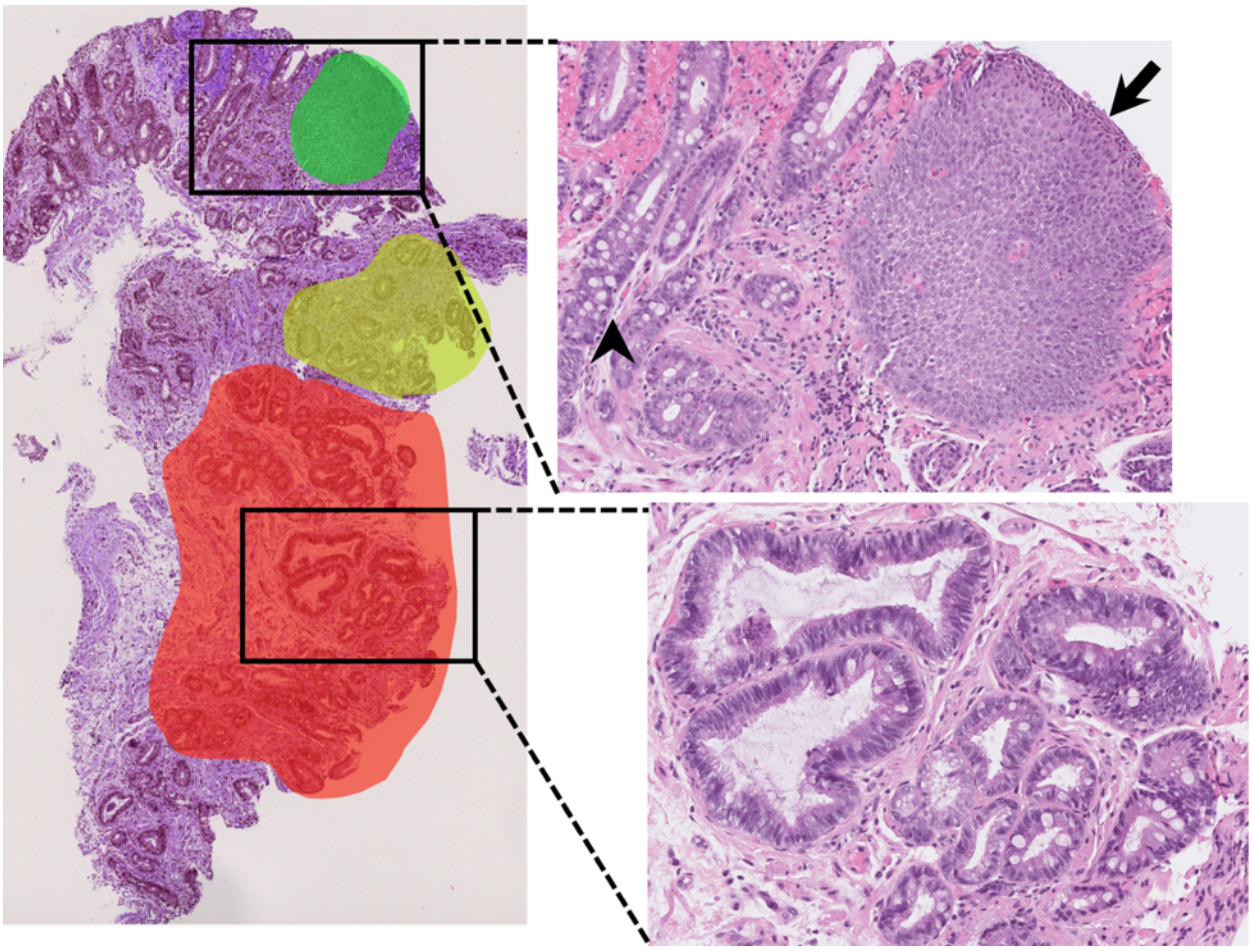

Figure 3.2: Example of the annotation process on a typical whole-slide image. Red, green, and yellow highlights indicate areas that were annotated and from which labeled patches were taken. Squamous tissue (black arrow), non-dysplastic Barrett's with Goblet cells (black arrowhead), and dysplastic tissue with crowding and hyperchromasia (lower zoomed section) were all present within the same whole-slide image.

K image augmentations (i.e. random flips and rescaling) $x_{i,k}$ or $u_{j,k}$, $k \in \{1, \ldots, K\}$. For an augmented input $x_{i,k}$ or $u_{j,k}$, the model, p, infers the probability output that the sample belongs to a certain class $p(x_{i,k})$ or $p(u_{j,k})$. To provide pseudo-labels for training, the model p guesses the label $p(u_{j,k}), \forall j$. To minimize entropy, MixMatch performs a "sharpening" procedure to accentuate the dominant class, such that:

$$q = \frac{1}{K} \sum_{k=1}^{K} p(u) \tag{3.1}$$

$$\hat{q} = \frac{q^{\frac{1}{T}}}{\sum_{l=1}^{L} (q_l)^{\frac{1}{T}}} \tag{3.2}$$

where L is the number of classes and $\hat{q}$ is the "sharpened" probability of a single unlabeled sample computed by averaging $p(u_j)$ of K augmentations. T is a hyperparameter to control the amount of accentuation/dampening–a lower T further sharpens the probability output.

To calculate the loss, MixMatch makes a distinction between samples originated from the labeled set and those from the unlabeled set. For samples in the labeled set $\mathcal{X}$, the loss is computed using a cross-entropy loss:

$$\mathcal{L}_{\mathcal{X}} = - \sum_{(x,y) \in \mathcal{X}} y * \log p(x) \tag{3.3}$$

where y is the ground truth binary indicator and p is the probability output for each class.

For unlabeled samples, $\mathcal{U}$, the loss is a mean squared error between the sharpened pseudo-label $\hat{q}$ and the probability output $p(u)$:

$$\mathcal{L}_{\mathcal{U}} = \sum_{u \in \mathcal{U}} ||\hat{q} - p(u)||^2 \tag{3.4}$$

Finally, the loss function then is the sum of the labeled loss and the unlabeled loss:

$$\mathcal{L} = \mathcal{L}_{\mathcal{X}} + \lambda_{\mathcal{U}} \mathcal{L}_{\mathcal{U}} \tag{3.5}$$

where $\lambda_{\mathcal{U}}$ is a weighting parameter. A higher $\lambda_{\mathcal{U}}$ would weigh inconsistencies found in the unlabeled dataset heavier than those in the labeled dataset. This unlabeled loss assumes that, for the most part, every sample belongs to the classes in-question.

MixUp

MixMatch makes use of training on generated synthetic data by using MixUp, which performs a pixel-level interpolation between images and pairwise interpolation between probability distributions. More formally, for a pair of two examples with their corresponding label probabilities $(x_1, p(x_1))$, $(x_2, p(x_2))$, I compute (x', p') by:

$$\lambda \sim \text{Beta}(\alpha, \alpha) \tag{3.6}$$

$$\lambda' = \max(\lambda, 1 - \lambda) \tag{3.7}$$

$$x' = \lambda' x_1 + (1 - \lambda') x_2 \tag{3.8}$$

$$p' = \lambda' p(x_1) + (1 - \lambda') p(x_2) \tag{3.9}$$

λ is a Beta distribution governed by the distribution's parameter α. MixUp is applied between any randomly chosen pairs of training samples, which could be either in the labeled set or unlabeled set. The synthetically generated samples of labeled, $\mathcal{X}'$, and unlabeled, $\mathcal{U}'$, data are then presented to the model for training.

In sum, the combination of the novel loss function and MixUp allows MixMatch to train on pseudo-labels to create an illusion of training on a fully annotated dataset.

FixMatch

FixMatch is a semi-supervised method has shown to perform well on semi-supervised benchmark datasets. Its aggressive transformation, pseudo-labeling, and treatment of unlabeled errors using cross-entropy loss allows it to achieve the highest performance in CIFAR-10, CIFAR-100 [46], and SVHN [67], relative to MixMatch.

Data Augmentation

For an unlabeled input u_j, FixMatch performs two types of augmentation: 1) strong augmentation and 2) weak augmentation. Weak augmentation performs flips and small rotations on the input data. Strong augmentation performs severe transformations like RandAugment [17] or CTAugment [9] on the input data. This work only applies RandAugment. RandAugment randomly selects transformation methods and determines the severity of that transformation. RandAugment can manipulate an input's contrast, rotation, color, brightness, sharpness, etc. Compared to weak augmentation, strong augmentation drastically changes the features within an image.

Pseudo-labeling and Loss Function

FixMatch trains on labeled samples using the typical empirical risk minimization procedures using backpropagation. For unlabeled samples, FixMatch predicts the softmax value of an unlabeled image under a weak transformation. If the weakly transformed input exceeds a threshold τ, then it is treated like a pseudo-label in the form of a one-hot vector, where the correct label is set to 1; otherwise, 0. Next, the model uses the same input to generate a softmax score of inputs under a strong transformation. This generated output is used as the prediction to backpropagate against the one-hot label of the weakly transformed input. If an unlabeled sample's softmax score is greater than τ, then it is treated just like a labeled

sample. The experiments in this study fixes $\tau = .95$, the default threshold used to produce the results in the original paper.

Comparison

This section compares and contrasts the two methods based on their respective pseudo-labeling, data augmentation, and unlabeled loss:

For pseudo-labeling, MixMatch infers labels by averaging the probabilities of various transformations applied to an unlabeled sample (e.g. simple horizontal and vertical rotations). This average probability score is then accentuated using a "sharpening" procedure where the score of the higher class probabilities increases and the scores of the lower class probabilities are dampened. The intuition is that if the model, on average, finds that a patch is of a certain class despite multiple transformations, then the best guess label of this patch is the class with the highest probability. Thus, sharpening this score increases the confidence that a patch belongs to a certain class, and it is used as the label for training. On the other hand, FixMatch uses the softmax output of the weakly augmented input. If the model infers a weakly augmented sample to have a softmax score greater than a predetermined threshold τ, then the model considers this softmax score as the pseudo-label. Unlike MixMatch which uses a sharpening proceedure to calculate a pseudo-label, FixMatch uses one-hot pseudo-labels.

For data augmentation, MixMatch applies a procedure called MixUp [94] on pairs of labeled or unlabeled samples to generate more synthetic data by performing a pixel-level interpolation between images and a pairwise interpolation between class probability distributions. This synthetic data, along with their interpolated pseudo-labels, are used for training the CNN. On the other hand, unlike MixMatch which uses only weak transformations, FixMatch uses both a strong and weak transformation on the unlabeled data point [9, 17].

Finally, for unlabeled loss, MixMatch uses the mean squared error (MSE) as the loss for the unlabeled samples. Compared to cross-entropy loss, MSE is less punitive to prediction errors. On the other hand, for FixMatch's unlabeled loss, the softmax output of the strongly augmented data point is compared against the one-hot encoded pseudo-label using a cross-entropy loss. If an unlabeled sample's softmax score is greater than τ, then FixMatch treats this sample equivalent to a labeled sample. Compared to MixMatch's MSE loss, cross-entropy loss severely punishes prediction errors.

Overall, FixMatch's relative aggressive pseudo-labeling, data augmentation, and unlabeled loss allows it to perform well in semi-supervised datasets. However, this work hypothesizes that these aggressive methods do not adapt well to the noisy setting within the biopsy domain.

Table 3.1: Class frequency (i.e. number of patches) of labeled, unlabeled, and test set.

Type	Dysplasia	Barrett's	Squamous	Total
Labeled Train	616	925	1,308	2,849
Unlabeled Train	-	-	-	889,028
Labeled Test	159	1365	1121	2,645

3.4 Barrett's Esophagus Data

A total of 387 slide images from 133 unique patients were collected. A selection of the whole-slide images were manually annotated to highlight examples of each class (squamous, Barrett's, and dysplasia) within each slide image (Figure 3.2).

To create the labeled dataset, from 29 of the total 387 slide images, 68, 51, and 85 segments of squamous, Barrett's, and dysplastic tissue were annotated, respectively. The segments were then subdivided into 1000x1000 pixel patches with 500-pixel overlap, and further curated to remove patches with excessive white space. All patches were extracted at the 40x magnification level. These clean samples were split at the patient-level into the labeled training dataset and testing dataset.

To create the noisy unlabeled dataset, the remaining slides were patched similar to the clean labeled set; however, no manual filtering was performed–leaving noise in the unlabeled dataset. The total training set contained 2,849 labeled patches and 889,028 unlabeled patches, and the test set contained 2,645 labeled patches (the model was blinded to these labels).

Table 3.1 summarizes the final class frequency of the dataset. Note the imbalanced nature of the dataset as the total number of dysplasia examples is disproportionately small due to its sparsity.

3.5 Experiments

This section compares and contrasts the classification performance of the two SSL methods under various label size conditions and imbalanced settings.

Implementation

All experiments use the ResNet-18 model. The default settings for both the MixMatch and FixMatch methods are used, except for FixMatch's learning rate which is set to $lr = .001$ (the default learning rate for MixMatch) as I have found it to converge better on the experimental dataset. For the unlabeled loss multiplier, I designated $\lambda_u = 1$ for both FixMatch and MixMatch. Both models are trained on 32 epochs with 512 iterations and a batch size of 22 samples. The reported implementations used 1024 epochs with a batch size of 64. However, I notice no increase in performance using 1024 epoch compared to 32 epochs. A Pytorch

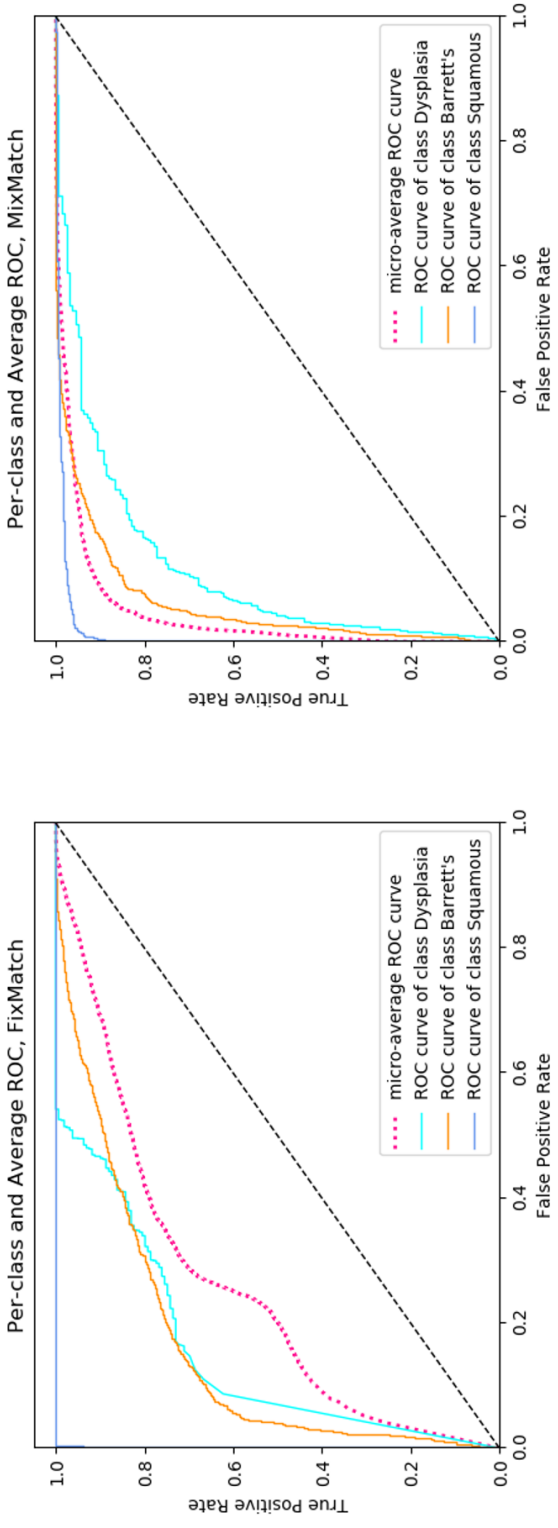

Figure 3.3: Per-class and average ROC curve of FixMatch and MixMatch trained on a (6, 12) patient-patch combination.

Table 3.2: Per-class and average AUC for the esophageal Barrett's dataset. Results show that MixMatch performs better than FixMatch on every (patient, patch) sampling levels. Notice, also, the improvement in performance as the number of patients increase.

		MixMatch			
Per-class patches	(Patient, Patch)	Dysplasia	Barrett's	Squamous	Micro-Ave.
36	(6, 6)	.91±.02	.97±.01	.99±.01	**.95±.01**
	(4, 9)	.89±.03	.96±.02	.99±.01	**.93±.03**
	(2, 18)	.83±.08	.92±.04	.98±.02	**.87±.04**
72	(6, 12)	.91±.01	.97±.01	.99±.01	**.95±.02**
	(4, 18)	.88±.05	.96±.01	.99±.01	**.95±.02**
	(2, 36)	.86±.03	.92±.05	.99±.01	**.90±.05**

		FixMatch			
Per-class patches	(Patient, Patch)	Dysplasia	Barrett's	Squamous	Micro-Ave.
36	(6, 6)	.82±.02	.74±.18	.99±.01	**.73±.12**
	(4, 9)	.83±.01	.88±.02	.99±.01	**.80±.09**
	(2, 18)	.79±.04	.80±.12	.99±.01	**.73±.10**
72	(6, 12)	.86±.04	.84±.13	.99±.01	**.81±.13**
	(4, 18)	.83±.04	.86±.06	.99±.01	**.77±.09**
	(2, 36)	.79±.04	.72±.16	.99±.01	**.70±.08**

	Full-Supervision			
	Dysplasia	Barrett's	Squamous	Micro-Ave.
All	.92±.01	.97±.01	.99±.01	**.96±.01**

implementation is used for both FixMatch[1] and MixMatch[2]. These implementations were verified to replicate their respective original results. Input data is resized from 1000x1000 pixels to 224x224 pixels and normalized between 0 and 1.

Performance comparison

The standard approach to analyze SSL methods is to measure their performance while varying the number of labeled samples. FixMatch and MixMatch are trained on two levels of labeled sample sizes: 36 and 72 patches per class; totaling 108 and 216, respectively. To test the effects of patient diversity, the number of patients from which are sampled is controlled. Six different patient-patch sampling combination levels are tested: (6, 6), (4, 9), (2, 18),

[1]https://github.com/valencebond/FixMatch_pytorch
[2]https://github.com/YU1ut/MixMatch-pytorch

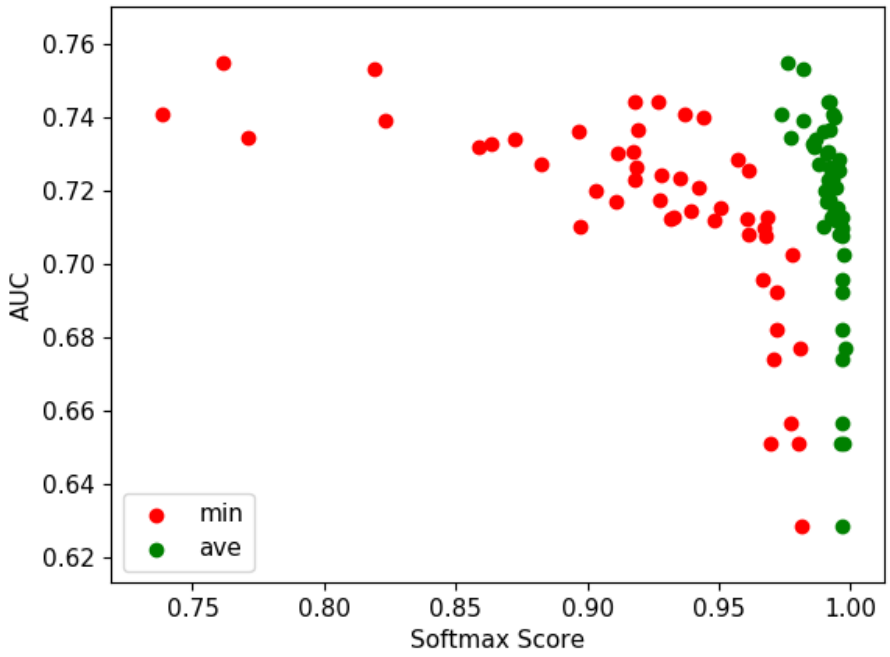

Figure 3.4: Effects of softmax scores of 10 open-set samples on the model's AUC for FixMatch using (2, 18) patient-patch combination.

(6, 12), (4, 18), and (2, 36). For example, the sampling level notation (4, 9) means that 4 random patients per class are sampled and, from each of these 4 patients, 9 random patches are sampled. These combinations were designed such that the number of total labels were held constant to control for labeled sample sizes. The average AUC and the per-class AUC are measured for each of these combinations over 5 trials. Table 3.2 shows that MixMatch performs better than FixMatch on the average AUC and dysplasia AUC. I also see that increasing the number of patients has a bigger impact on the performance of both the models compared to just increasing the number of labeled patches, signifying that patient diversity has a larger role on the performance of these models. Figure 3.3 compares the performance of an identical (6, 6) patient-patch sampling on both the FixMatch and MixMatch methods.

As a proxy to an upper bound, the model was trained using a fully supervised method trained on all the labeled samples. The fully supervised model's performance is comparable to MixMatch's performance at the (6, 6) and (6, 12) combination levels, despite MixMatch only having a small fraction of the total labels.

To offer an explanation as to why FixMatch produces poor results on the esophageal dataset, I designed a follow-up experiment by tracking the effects of the softmax score of 10 hand picked, open-set examples on the AUC using the (2, 18) combination. The model

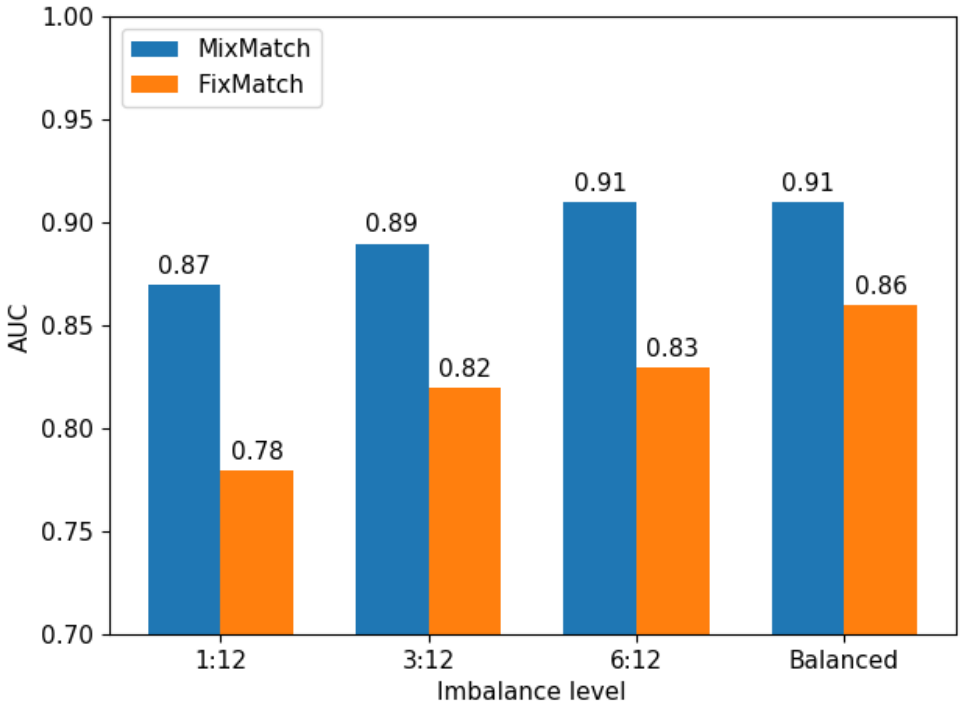

Figure 3.5: Dysplasia class's AUC score for varying levels of imbalances applied to a (6, 12) patient-patch combination. The various imbalance levels against the dysplasia class are 1:12, 3:12, 6:12, and balanced. The imbalance level 1:12 means that the dysplasia class will have 1 sample for 6 patients. And the Barrett's and squamous class will have 12 samples from 6 patients each.

AUC is measured at every 126 iterations for 64 cycles and the corresponding softmax scores that the 10 open-set examples produce. The softmax score is the probability that a given sample belongs to a certain class. Figure 3.4 shows the minimum and average softmax scores of the last 48 cycles. This shows that, as the model erroneously becomes more confident of the open-set examples, the model's performance deteriorates as well. More interestingly, the model begins to deteriorate when the minimum softmax score exceeds .95 (FixMatch's default threshold value).

Effects of imbalances

To test the effects of the various degrees of imbalances, the patient-patch sampling combination is fixed at (6, 12). Then the amount of labeled dysplasia samples are decreased to 1, 3, and 6 samples per patient–totaling 6, 18, and 36 samples for the dysplasia class, respectively, compared to the 72 samples for Barrett's and squamous. The average AUC

is measured across 5 trails. Figure 3.5 shows that MixMatch is more robust to imbalances compared to FixMatch on average. More interestingly, MixMatch has a higher average AUC on the imbalanced dysplasia class, and comparable to the balanced result at (6, 12) combination. Overall, however, both methods degrade with high level of imbalances due to their performance on the dysplasia (minority) class.

3.6 Discussion

This study shows that, with only a few exemplary images, one can train an effective model to detect esophageal disease patterns on histopathology datasets. In this study, I compare and contrast the performance of MixMatch and FixMatch. Although FixMatch performs better overall in general computer vision datasets, our results show that MixMatch performs better on histology datasets–where noisy, open-set samples are present. Also, MixMatch's pseudo-labeling and data augmentation procedures are more robust to the impact of histology datasets, even under varying degrees of imbalanced scenarios. Finally, our experiments show that patient diversity has a significant impact on the performance of SSL methods.

While there could be many compounding factors as to why FixMatch performs poorly on datasets with open-set samples, I hypothesize that one major reason for FixMatch's poor performance is due to its use of the thresholding method for pseudo-labeling and cross entropy loss to account for errors: the thresholding method incorrectly labels an open-set sample as one of the classes in-question and the cross-entropy loss impels the model to over-confidently predict an open-set sample as belonging to one of the classes in-question.

3.7 Chapter Summary

This work contributes to the body of literature pertaining to SSL in medical imaging. Applying the leading SSL methods to the problem of detecting disease in histology images, this study concludes the MixMatch performs better than FixMatch. This work motivates the development of SSL methods that are robust to open-set noise common in histology datasets. In the next chapter, I introduce a two-model architecture–CoMixMatch–that improves the performance of semi-supervised methods on noisy biopsy datasets. In the next chapter, I propose a solution that improves MixMatch's performance against noisy, larger-scale datasets.

Chapter 4

CoMixMatch: Semi-supervised Classification of Noisy, Gigapixel Histology Images

4.1 Overview

In order to circumvent the laborious annotation process, some researchers have turned to semi-supervised learning techniques where models learn from a large body of unlabeled data along with a smaller set of labeled data. However, these techniques have not been fully examined in the histology setting where there is a high degree of noise. This chapter investigates an extension of the semi-supervised method MixMatch–CoMixMatch–which applies semi-supervised co-teaching and a contrastive unlabeled loss. More specifically, I study these models' impact under a highly noisy, open-set histology setting. The findings here motivate the development of semi-supervised methods to ameliorate annotation costs commonly encountered in medical data applications.

4.2 Background

Convolutional Neural Networks (CNNs) have demonstrated to be a powerful solution to many computer vision challenges. While the computing resources needed to train and store large-scale CNNs have become increasingly cheaper and more democratized as the barrier to train custom deep neural networks is lowered, some of the larger costs today are from activities relating to the annotation of datasets for training and evaluating these models. These costs are exacerbated in the field of medicine where physicians' time is costly–presenting a high obstacle to apply traditional machine learning techniques on medical data.

Thus, researchers turn to techniques like semi-supervised learning (SSL) to minimize the laborious, annotation process required by fully-supervised methods. SSL techniques adapt to a setting where one has access to a small amount of labeled data and a larger proportion

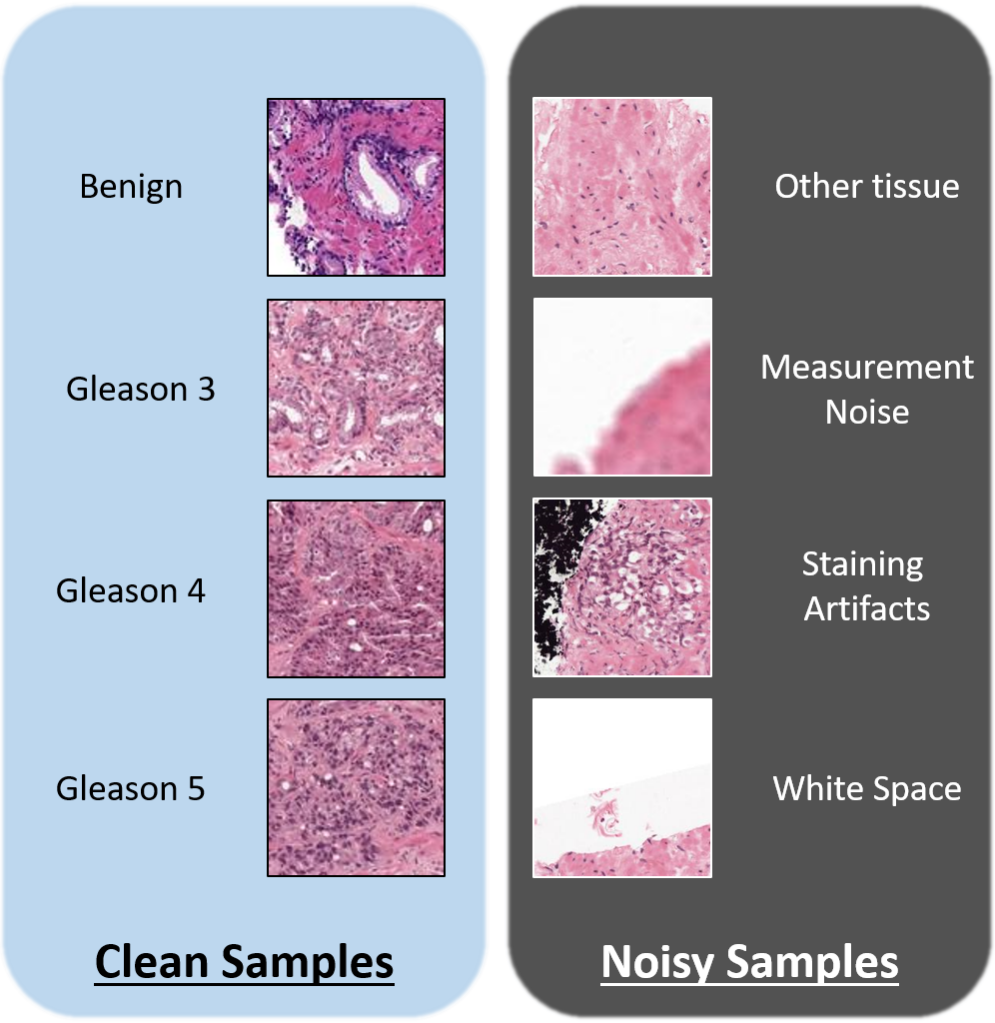

Figure 4.1: Distinction between clean and noisy patches in the PANDA dataset whose goal is to grade lesions into 4 categories (in order of severity: Benign, Gleason 3, Gleason 4, and Gleason 5). Noisy patches found in the unlabeled training set may introduce bias by having the model overfit to the wrong patterns.

of unlabeled data. Recently, there has been a surge in the state-of-the-art performance of semi-supervised learning with methods like MixMatch [8] outperforming other previous SSL methods on benchmark computer vision datasets.

Although MixMatch and other recent techniques are empirically successful in general SSL tasks, they have not been thoroughly studied under conditions common in the field of histology. This body of work contributes to the growing literature of machine learning in medicine by introducing CoMixMatch, a semi-supervised technique extending MixMatch.

CoMixMatch is shown to be more robust to the noisy conditions in histology datasets. CoMixMatch improves from its predecessor by 1) applying a two-model SSL co-teaching method, and 2) replacing the unlabeled loss with a two-model contrastive loss.

Digitized biopsy slides are high resolution images much larger than standard images. These high resolution sizes are prohibitive for the application of neural networks as 1) resizing these images would destroy the patterns important to the diagnosis, and 2), if resizing was not performed, off-the-shelf GPUs do not carry enough memory to store the parameters of a large model required by a gigapixel-sized input.

To circumvent this problem, researchers perform "patching" operations by subdividing the slide images into smaller patches–cropped in a sliding window manner–and use them as input data to the CNN. These patches should be small enough to fit into GPU memory and carry enough detail to discern if disease is present. Patching has performed well on disease detection on biopsy slides [33, 90, 86, 16, 13, 95].

While fully-supervised methods have shown successes, SSL techniques have not been fully studied as it applies to histology challenges. In the SSL setting, the model is privy to only a few labeled patches, and has access to a larger set of unlabeled patches. Because users have no prior knowledge of where the relevant areas are located in the unlabeled slides, they cannot filter noisy patches that may carry open-set patterns that are not relevant to the context of the problem. These open-set areas could be caused by tissue outside the domain of inquiry, sensor noise, imperfections in the staining process, white space, etc. Figure 4.1 shows some examples of open-set noise as compared to clean samples. In the case where a high number of open-set patches is present in the training data, CNNs will inadvertently overfit to these images [93] and may learn the wrong patterns; thus, degrading the generalizability of the model.

4.3 Methods

For this section, I introduce CoMixMatch–a method that adapts to the open-set SSL setting. I will then detail the pre-processing step of the experimental data. For this study, I use the PANDA dataset–a dataset consisting of prostate cancer biopsies–as the use case. MixMatch is also used in the experiment; however, the reader may find a review of the MixMatch method in Section 3.3.

CoMixMatch

This section describes CoMixMatch to improve SSL detection for diseases in biopsies. In this section, I introduce 1) SSL co-teaching, which aims to avoid having the model overfit to the noisy samples, and 2) the unlabeled contrastive loss replaces MixMatch's unlabeled loss, $\mathcal{L}_{\mathcal{U}}$.

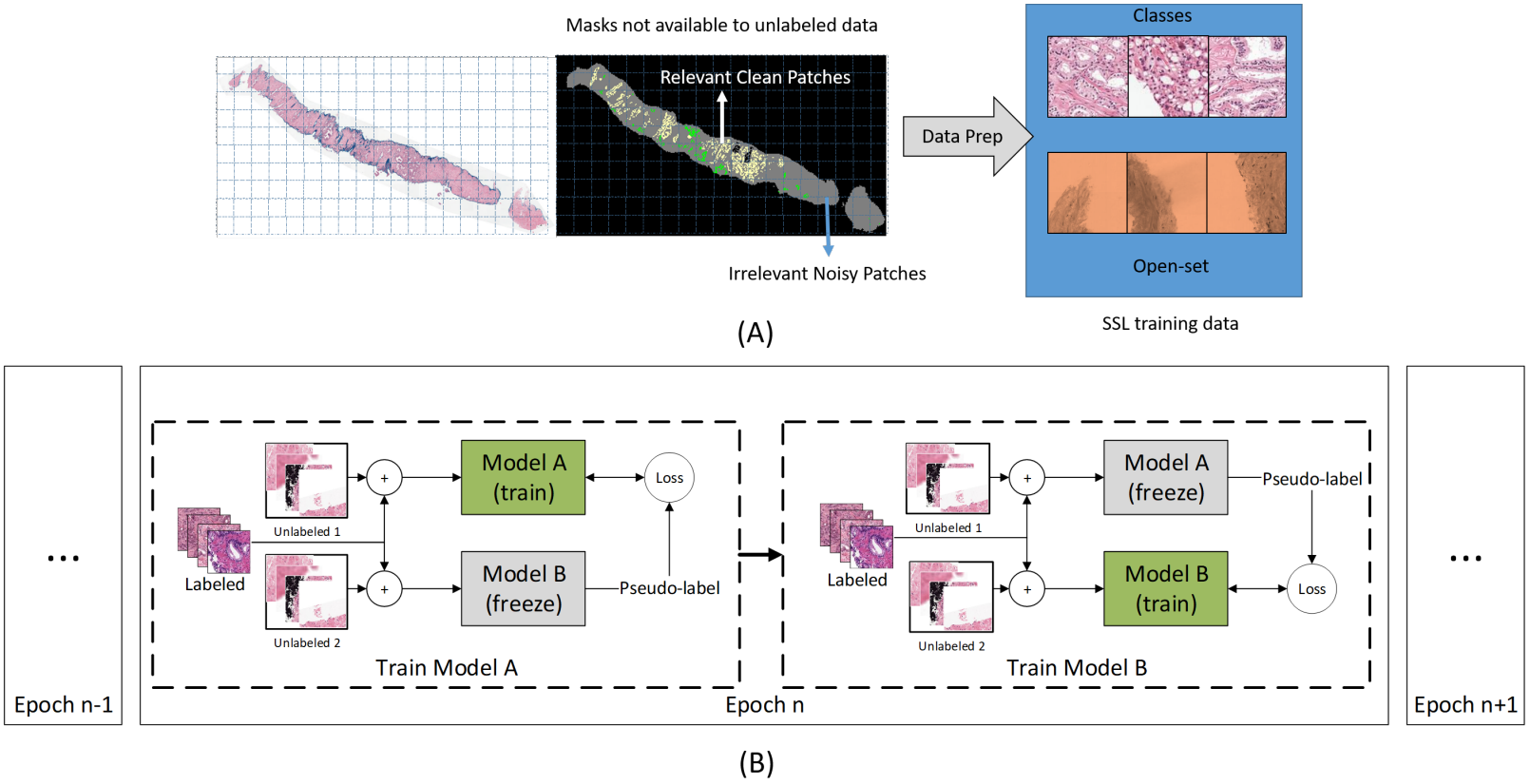

Figure 4.2: (A) Depicts the pre-processing step of subdividing biopsies into patches that will be fed into a CNN. Due to the giga-pixel sized image, the common practice is to "patch" the image in a sliding window manner. In the SSL setting, users have access to a large body of unlabeled data and only a small amount of labeled data. During the patching process, I cannot discern if a certain patch is relevant to the problem; thus, in the biopsy case, SSL techniques should include a capability to handle noisy open-set data.
(B) Diagram of the CoMixMatch training process: for each epoch, Model B is frozen while Model A is trained. I use Model B's pseudo-labels to avoid having Model A overfit to the open-set noise. Once Model A reaches the end of its training phase, Model B is then trained using the pseudo-labels of Model A. Each model is only trained on a non-overlapping set of the unlabeled dataset.

Algorithm 1 CoMixMatch

Input: co-teaching models p_1 and p_2, labeled dataset $\mathcal{X}$, unlabeled dataset $\mathcal{U} = \mathcal{U}_1 \cup \mathcal{U}_2$, unlabeled loss weight $\lambda_{\mathcal{U}}$, Beta distribution parameter α, number of augmentations K

Alternate training p_1 and p_2 for every epoch

train p_1, freeze p_2
for Every iteration step **do**
    From $\mathcal{X}$, draw a mini-batch $\{(x_b, y_b); b \in \{1, \ldots, B\}\}$
    From $\mathcal{U}$, draw a mini-batch $\{u_b; b \in \{1, \ldots, B\}\}$
    $q_{1,k} = p_1(u_k)$ and $q_{2,k} = p_2(u_k)$
    $q_{ave} = \text{Average}(q_{1,k}, q_{2,k})$
    $\mathcal{U}_1' = \text{Concatenate}(\mathcal{U}_1, q_{ave})$ (pair samples and pseudo-labels)
    $\hat{\mathcal{X}}, \hat{\mathcal{U}}_1' = \text{MixUp}(\mathcal{X}, \mathcal{U}_1', \alpha)$
    Calculate $\mathcal{L}_{\mathcal{X}}(\hat{\mathcal{X}}, p_1(x))$
    Calculate $\mathcal{L}_{\mathcal{U}}(\hat{\mathcal{U}}_1, q_{1,k}, q_{2,k})$ # Contrastive Loss: euclidean distance of the output of two models
    Backpropagate with $\mathcal{L}^{(1)} = \mathcal{L}_{\mathcal{X}}^{(1)} + \lambda_U \mathcal{L}_{\mathcal{U}}^{(1)}$
end for
train p_2, freeze p_1

Semi-supervised co-teaching

Co-teaching (Han et. al.) [27] has been used primarily to address the issue of noisy labels in the fully-supervised setting. Its purpose is to have models robustly learn despite samples being incorrectly labeled (i.e. labels are flipped). The authors have found that models first learn the patterns of the clean data then, overtime, it overfits to noisy data. To prevent overfitting, Han et. al. introduced co-teaching, which simultaneously trains two networks while each network helps the other to filter errors.

In the SSL setting, when models overfit to the noisy unlabeled samples, they accumulate confirmation bias through their pseudo-labeling–thus the model accumulate errors progressively. CoMixMatch adapts the ideas in Han et al. to the semi-supervised setting by using a two-model method, p_1 and p_2, that alternate training sequentially. Figure 4.2(B) summarizes the co-teaching training process. When p_1 is in its training phase, p_2's parameters are kept frozen. The labeled loss function, $\mathcal{L}_{\mathcal{X}}^{(1)}$, used to backpropagate when training p_1 and p_2 is frozen is:

$$\mathcal{L}_{\mathcal{X}}^{(1)} = -\sum_{(x,y)\in\mathcal{X}} y * \log p_1(x) \tag{4.1}$$

At test time, CoMixMatch's prediction, $\hat{y}$ is the average output of the two models:

$$\hat{y} = \frac{p_1(x) + p_2(x)}{2} \tag{4.2}$$

		CoMixMatch ($\lambda_u = 1$)				
	(Patients, Patches)	Benign	Gleason 3	Gleason 4	Gleason 5	**Macro Ave.**
$\lambda_u = 1$	(4, 6)	.99±.01	.95±.02	.89±.04	.90±.11	**.94±.02**
	(3, 8)	.99±.01	.95±.02	.84±.01	.95±.01	**.93±.01**
	(2, 12)	.99±.01	.92±.04	.81±.14	.76±.05	**.87±.03**
		MixMatch ($\lambda_u = 1$)				
	(Patients, Patches)	Benign	Gleason 3	Gleason 4	Gleason 5	**Macro Ave.**
	(4, 6)	.98±.01	.95±.01	.86±.05	.90±.07	**.92±.01**
	(3, 8)	.98±.01	.95±.02	.73±.06	.91±.04	**.89±.01**
	(2, 12)	.90±.14	.92±.03	.65±.15	.68±.19	**.79±.03**

		CoMixMatch ($\lambda_u = 75$)				
	(Patients, Patches)	Benign	Gleason 3	Gleason 4	Gleason 5	**Macro Ave.**
$\lambda_u = 75$	(4, 6)	.99±.01	.96±.01	.90±.02	.90±.09	**.94±.02**
	(3, 8)	.99±.01	.95±.02	.83±.05	.94±.01	**.93±.02**
	(2, 12)	.98±.02	.93±.03	.79±.15	.72±.12	**.86±.04**
		MixMatch ($\lambda_u = 75$)				
	(Patients, Patches)	Benign	Gleason 3	Gleason 4	Gleason 5	**Macro Ave.**
	(4, 6)	.97±.02	.93±.04	.84±.06	.84±.10	**.89±.05**
	(3, 8)	.96±.02	.92±.05	.79±.11	.90±.01	**.89±.04**
	(2, 12)	.93±.05	.94±.02	.69±.13	.70±.17	**.83±.08**

	Fully-supervised				
	Benign	Gleason 3	Gleason 4	Gleason 5	**Macro Ave.**
All	.99	.99	.98	.97	**.99**

Table 4.1: SSL and fully-supervised AUC (mean±std) comparison on 1) two different levels of λ_u, the hyperparameter that controls the learning sensitivity to the unlabeled loss, and 2) three different (patient, patch) sampling configuration: (4,6), (3,8), (2,12); each totaling 96 labeled samples. The fully-supervised model is trained on the patches of 564 patients. The SSL results were averaged over three trials.

Unlabeled contrastive loss

Unlike MixMatch, whose unlabeled loss $\mathcal{L}_{\mathcal{U}}$ calculates the mean squared error between a pseudo-label and its "sharpened" label, CoMixMatch uses a contrastive loss measuring the

distance of the outputs of the paired models p_1 and p_2:

$$\mathcal{L}_{\mathcal{U}}^{(1)} = \sum_{u \in \mathcal{U}} ||p_1(u) - p_2(u)||^2 \tag{4.3}$$

where $p_m(u)$, $m \in \{1, 2\}$ is the average output of k augmentations of input u. This contrastive loss prioritizes agreement between models while de-emphasizing the need to produce a 'forced' probability class output. Finally, note that I do not introduce any sharpening in the pseudo-labels as it introduces bias to the model. The total loss for model p_1 is:

$$\mathcal{L}^{(1)} = \mathcal{L}_{\mathcal{X}}^{(1)} + \lambda_U \mathcal{L}_{\mathcal{U}}^{(1)} \tag{4.4}$$

Once model p_1 is finished training, we freeze p_1 parameters and train p_2.

4.4 Gleason Grading Data

To evaluate the model, I use the Prostate cANcer graDe Assessment (PANDA) [1] [82] dataset which is a dataset comprised of prostate biopsies. The main challenge in this competition is to classify the severity of prostate cancer from microscopy scans of prostate biopsy samples. The dataset grades cancerous areas of tissue into 5 categories: Background, Stroma, Benign, Gleason 3, Gleason 4, Gleason 5. This dataset contains both images and masks which segments areas of the disease (See Figure 4.2(A)). I treat Benign and Gleason 3-5 label as our classes in-question. Stroma label acts as the noise in the dataset. I discard the Background portions of the dataset as they are relatively easy to detect using color thresholding methods.

To simulate an SSL environment, I divide the dataset into the labeled (clean) and the unlabeled (noisy) dataset: 706 patients in the labeled dataset and 1,985 patients in the unlabeled dataset.

To create the noisy dataset, slides are subdivided into patches in a slideing window manner with a window size of 512x512. Data with high amounts of white space are discarded. Each sample is patched in such a way that it is agnostic to the presence of diseases in-question, and may or may not contain disease features.

To simulate a human annotated clean dataset, the PANDA segmentation mask is used to provide the location of the disease in the biopsy. The biopsies are sampled in areas where disease is present (Benign and Gleason 3-5). A patch is only admitted into the clean dataset if at least 50% of the area is of a certain class. The patch selection is designed such that no other disease is present other than the disease labeled (i.e. there are no multi-class patches). Table 4.2 shows the final per-class patch count of the dataset used.

4.5 Results

This section compares the performance of CoMixMatch, MixMatch and fully-supervised models.

[1] https://www.kaggle.com/c/prostate-cancer-grade-assessment/data

Table 4.2: Class frequency of clean and noisy patches. This study uses two types of data: clean and noisy. The clean dataset simulates data that were manually labeled by humans while the noisy dataset simulates the un-manicured data in the unlabeled dataset. The samples in the clean and noisy datasets are split at the patient-level, i.e. no one patient can be found in both clean and noisy samples.

Type	**Benign**	**Gleason 3**	**Gleason 4**	**Gleason 5**	**Total**
Labeled	464,334	42,343	7,085	2,112	99,823
Unlabeled	-	-	-	-	864,334
Test	9,324	676	1,626	360	11,986

Implementation

I implement CoMixMatch and MixMatch using a ResNet-34 architecture [30]. If not otherwise mentioned, the default settings are used from the original paper. I train both models on 48 epochs with 1024 training steps per epoch with a batch size of 64. Because CoMixMatch has two distinct models, each model will be trained only for 512 iterations per epoch to give a fair comparison against the MixMatch baseline. While the original reported implementation used 1024 epochs with batch size of 64, no increase in performance is observed using 1024 epochs compared to 48 epochs. Both models used a Pytorch implementation [2] and are verified to approximately replicate the reported results. Input images are resized from 512x512 pixels to 112x112 pixels. Pixel values are normalized between 0 and 1. Each model is trained on a Titan V GPU for approximately 7 hours. For all comparison studies, the per-class AUC and class-normalized average (macro-) AUC are used as a standard metric to compare the overall performance of models. The SSL models have access to the entire unlabeled dataset but only a portion of the clean dataset. The benchmark fully-supervised models have access to the entire labeled dataset (unless otherwise stated). All models were validated on a test set using a random subsection of the clean dataset comprising of 141 patients randomly selected at each trial. All models were blinded to the test set during training.

Relative Performance

This section tests the classification accuracy of CoMixMatch against MixMatch and fully-supervised methods. To test the impact of the labeled sample size, I designed a sampling process that fixes the number of labeled samples for each trial. I choose a (m, n)-sampling protocol where I select m amount of patient for each class, and, from each sampled patient, I select n patches. The models are tested at 3 different levels $(4, 6)$, $(3, 8)$, and $(2, 12)$–totaling 96 patches. For example, at the $(4, 6)$ sampling level, for each class, I randomly select 4 patients, and, for each of these patients, I select 6 patches. This manner of sampling allows us to study the effects of patient diversity. The central assumption is that patches from the

[2] https://github.com/YU1ut/MixMatch-pytorch

same patients would contain the same patterns, and, as I decrease the number of patients sampled, the variance of the labeled dataset decreases.

Table 4.1 shows that increasing the number of patients improves the performance of both SSL methods. Moreover, CoMixMatch performs better than MixMatch at all (m, n)-sampling levels. I also observe that CoMixMatch is more resilient compared to MixMatch as the number of patients decrease.

To test the sensitivity of the models on the open-set noise in the dataset, I evaluated the model on on two unlabeled loss values: $\lambda_{\mathcal{U}} = \{1, 75\}$. A higher $\lambda_{\mathcal{U}}$ would shift the models to prioritize organizing the unlabeled dataset according to their respective loss function. A degradation in performance would indicate a method that is sensitive to noise in the unlabeled dataset. In Table 4.1, between the two $\lambda_{\mathcal{U}}$ levels, I see a decrease in the average performance of both models indicating the negative impacts of openset noise in the unlabeled dataset. However, MixMatch significantly decreases in performance more so than CoMixMatch, indicating that CoMixMatch is more robust to open-set noise relative to MixMatch. Note that the fully-supervised model, with access to all 564 labeled patients, still has a better performance than both SSL methods.

To test the additional benefit of using a large unlabeled dataset, I evaluate CoMixMatch against a fully-supervised method at different patients-per-class levels. To design a scenario that mirrors real-world application, for this experiment, I do not restrict the number of patches. I trained CoMixMatch and fully-supervised ResNet-34 on four different levels of patients-per-class: 4, 8, 12, 16. Figure 4.3 shows that CoMixMatch outperforms the fully-supervised case at all levels. While this is expected, note that CoMixMatch's macro-average AUC at 4 patients per-class (.97) is comparable to that of the fully-supervised macro-average AUC at 16 patients per-class (.95)–which indicates that CoMixMatch only requires a quarter (4 out of 16 patients) of the annotations for this particular dataset. Lastly, at the 16 patient-per-class (64 total) level, CoMixMatch approaches the performance of the fully-supervised approach using all 564 labled patients.

4.6 Chapter Summary

This body of work introduced CoMixMatch–a method which extends MixMatch to provide more robustness against open-set noise in the SSL setting. I show that CoMixMatch performance improves upon MixMatch and fully-supervised methods on low amounts of labeled data, so long as it has access to a larger amount of unlabeled data. Although the fully-supervised method, with access to all labeled dataset, still outperforms the SSL methods at every level, I achieve somewhat comparable results at a lower annotation cost.

Future work includes the continuation of developing techniques designed to decrease the cost of building computer-aided diagnostic tools by pairing CoMixMatch with active learning approaches. Furthermore, while CoMixMatch has shown promising performance on open-set medical datasets, I conjecture that the techniques presented are general enough to be more broadly applied to other computer vision datasets.

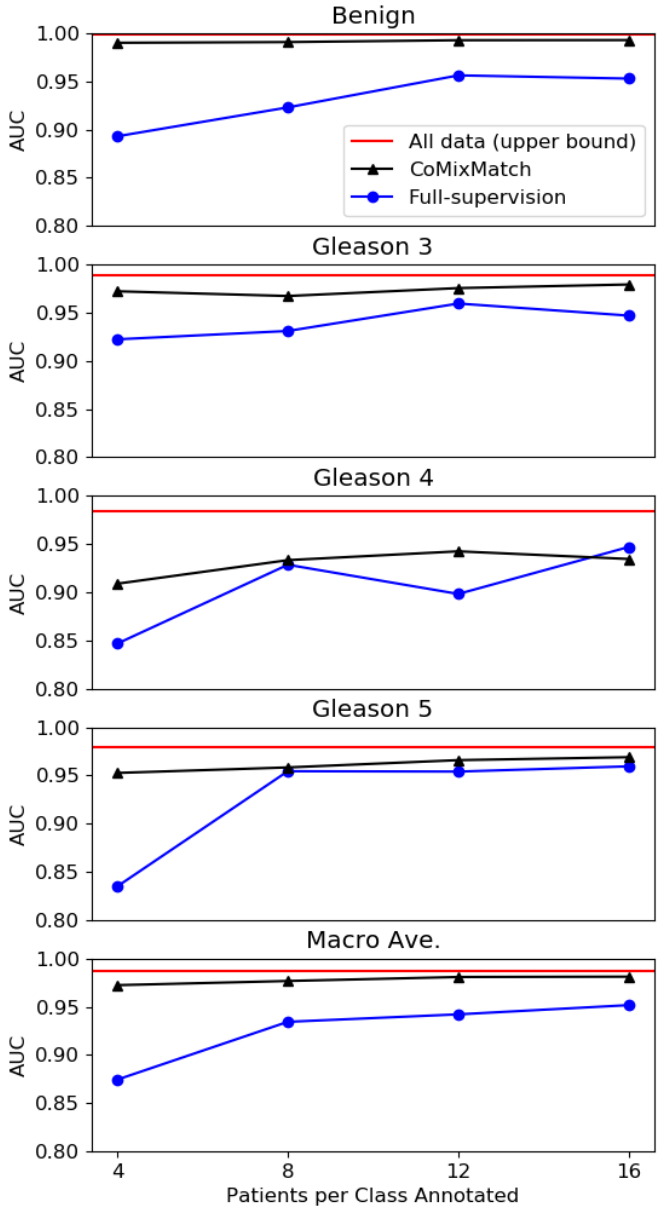

Figure 4.3: Average AUC performance comparison of CoMixMatch against a fully-supervised method at various number of annotated patient over three trials. Observe that CoMixMatch outperforms the fully-supervised method with access to the same labeled data. All data (upper bound) is a fully-supervised model trained on 564 patients.

Chapter 5

Class Imbalances in Biopsy Images with Polarized Features

5.1 Overview

Imbalanced classes often occur in many medical data challenges. When dealing with these types of data conditions, deep learning methods perform poorly as the learned patterns are biased towards those in the majority classes. Current methods that address class imbalance involve two-stage methods that, first, learn features from the majority class, then transfer these learned features to the task of classifying all classes. These methods assume that features in the majority class are rich enough to differentiate across the minority class. This chapter argues that this assumption does not hold in some biopsy cases, where cancerous features are found only at the tail-end of the cancer progression. This polarization of features causes sub-optimal performance when applying current imbalanced methods. Furthermore, I discover that there are some classes, possibly outside the majority class, that contain robust features which, if detected and leveraged, may alleviate the problems caused by class imbalances.

5.2 Background

CNNs have demonstrated to be powerful solutions to many computer vision challenges. Nevertheless, they still succumb to problems commonly found in the real-world. Typical training schemes usually assume that class samples are plentiful and well-balanced. In most realistic cases, however, scarce data is a common constraint in developing these powerful tools. This is exacerbated in the field of medicine where certain diseases are rarely encountered causing severe imbalances in training datasets where a few classes hold the majority of the samples while a few classes hold a minority. This data condition causes poor performances when training these models using traditional techniques, as the model is biased towards learn-

ing the patterns of the majority class; thus, presenting a high obstacle to apply traditional methods on medical data.

To alleviate the effects of imbalances, researchers have proposed effective methods that aim to solve the problem of imbalanced datasets using two-stage technique [38]. In the first stage, a model learns features from the instance rich majority classes–producing a robust feature extractor. In the second stage, the convolutional layers are frozen and the classifier is re-trained using class-balanced batches. This paradigm, of learning features from the majority to classify the minority, has proven to be successful on many general computer vision benchmark datasets.

However, it is yet unclear if this paradigm holds in cases where the majority classes do not contain the proper features necessary to distinguish between the minority classes. Biopsy classification problems exist under this polarized case where cancer patterns exist at the terminal disease classes. At the same time, scarcity of samples increases as classes move from normal to disease as well (See Figure 5.1). When using two-stage methods, cancer makers are not abundantly available to learn a robust feature model, causing sub-optimal performance to classify rare diseases.

Thus, one can ask if there exists alternative classes outside the set of majority classes that contain robust features that allows the proper discrimination across all classes. Detecting and learning features from these classes could alleviate the problems of imbalances by either 1) collecting more of these classes or 2) developing methods that promote the feature learning of these classes. This work aims to empirically show that some biopsy problems have this property. For this study, I will use the PANDA dataset biopsies as our use case.

This body of work contributes to the growing literature pertaining to the training of deep models under class imbalanced scenarios. Our goal is to alleviate the effects of imbalanced problems common in medical machine learning by analyzing the effects of polarized features in the imbalanced scenario. First, I show that, in the case of the Gleason grading system, polarized features affect the performance of the model. Secondly, using this knowledge, I introduce the idea that leveraging alternative classes containing robust global features could provide ways to alleviate the effects of imbalanced problems under resource-constrained environments.

5.3 Related Works

In this section, I introduce the current methods to tackle the problem of class imbalance and introduce the Gleason cancer grading system as this study's use case.

Class Imbalances

Imbalanced datasets have been studied for decades. Typical solutions have been to perform data re-balancing methods like undersampling [85, 48, 64, 80], and oversampling [12, 28, 29].

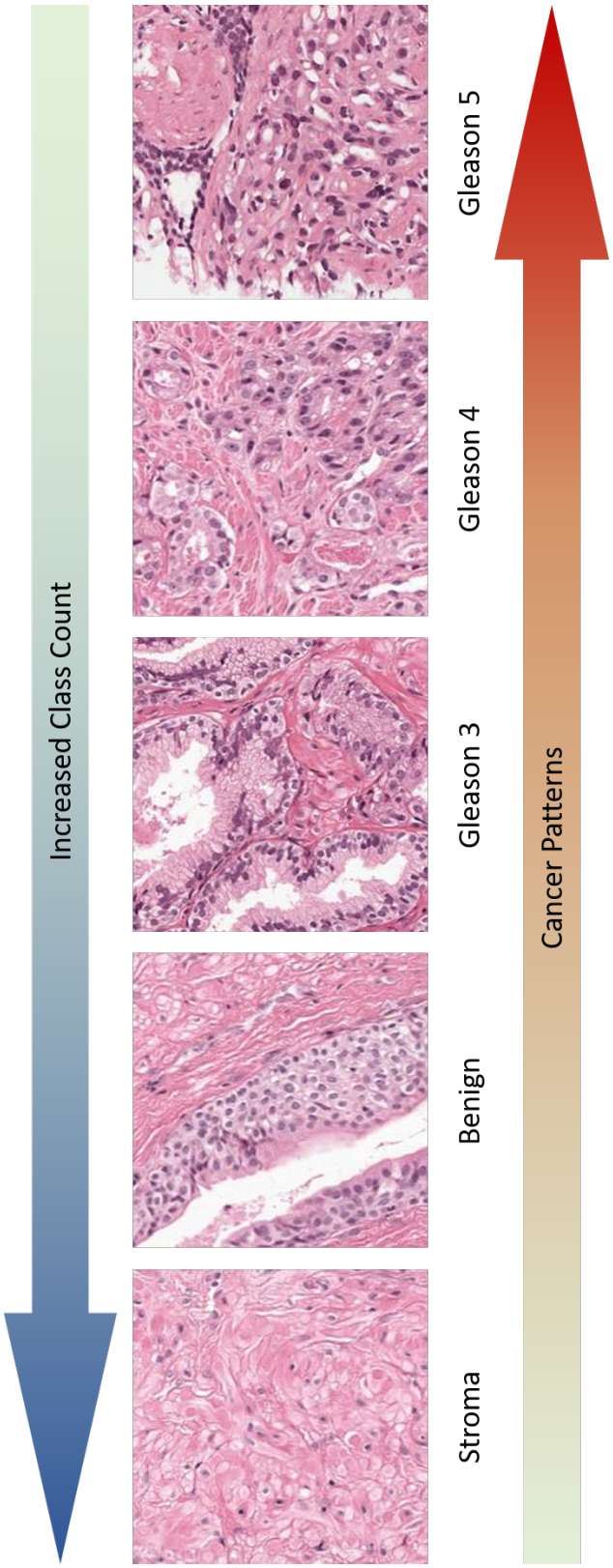

Figure 5.1: Diagram depicting the sequential progression of the Gleason patterns in order by severity (from left to right). Simultaneously, as cancer patterns increase, class counts decrease. Benign glands have pale cytoplasm with having small and regularly shaped nuclei. The glands are grouped together. Gleason Pattern 3 has no loss of glandular differentiation. Gleason Pattern 4 has partial loss of glandular differentiation. Gleason Pattern 5 has an almost complete loss of glandular differentiation.

In the deep learning era, there have been exciting advances proposing solutions to the problem of imbalances that scale to datasets that mirror the expanse of the real world. These long-tailed datasets [59] have classes that follow a Zipfian distribution where a minority of the classes (called the head) hold a majority of the samples and a majority of the classes (called the tail) hold a minority of the samples.

To tackle this problem, Kang et al. [38] proposed an effective method, called Decoupling, that follows a two-stage approach by detaching the process of learning features and learning discriminators. This successful paradigm for tackling imbalances has been extended to many other techniques [83, 96, 83, 57].

This study directly relates to these works aiming to alleviate the burden of imbalanced datasets as it applies to biopsy datasets. Whereas these works applied their techniques on seemingly arbitrary head classes, I examine these methods under a case with polarized features.

Gleason Grading System

Although extremely deadly when left untreated, prostate cancer patients can improve their survivability if the disease is detected early. To treat this disease, pathologists examine biopsy samples to determine the severity of the disease and their accompanying treatment [20]. Pathologists classify areas into 5 classes: Stroma (S), Benign (B), Gleason 3 (G3), Gleason 4 (G4), and Gleason 5 (G5), by order of severity. These classes, to some degree, follow a long-tailed distribution, where S is the most abundant and G5 is the rarest.

Pathologists examine these digitized biopsies as gigapixel-sized images. These high resolution sizes are prohibitive for the application of neural networks as resizing these images would destroy microscopic patterns important to the diagnosis, and, if resizing was not performed, off-the-shelf GPUs do not carry enough memory to store the parameters of a large model required by a large input. To ameliorate the problem, the common practice is to perform "patching" operations by subdividing the slides into smaller patches. These patches are used as input data to the CNN. In the next section, I will detail the pre-process patching method used in our experiments.

5.4 Methods

In this section, I summarize the idea of decoupling the representation and classification task to address class imbalances. I then expound on its implications to address class imbalances where features are polarized to the tail. Finally, I discuss the pre-processing of the PANDA dataset which is the large-scale dataset I used to perform the experiments.

Decoupling

Kang et al. [38] introduced a simple yet effective two-stage method to address class imbalances. First, a model is trained while uniformly sampling the training set. Then, the convolutional base layers are frozen while a re-initialized classifier is re-trained using class-balanced batches. Kang et al. also introduced parameter norm adjustment methods to slightly improve the classification performance. In this study, I will only use the standard classifier retraining (cRT) without parameter adjustments as I observed it to perform the best on our dataset.

The main idea behind decoupling is that the model first learns how to organize the feature-rich head and, subsequently, transfer these features for the purpose of classifying the tail. This idea assumes that the model can learn robust inter-class features from the head classes and transfer these features to differentiate the tail classes. More specifically, keeping with Tang et. al.'s [83] analogy, in order to classify a tail class (e.g. "remote"), it would be useful for the model to learn the features of the head classes (e.g. "couch" or "TV") associated with such a tail-class. However, if the head classes only contain non-associated features (e.g. "swimming pool" or "ocean"), in theory, the tail class should be poorly discernible.

As mentioned previously, in the general biopsy case, I hypothesize that disease markers only manifest themselves fully in the tail classes (See Figure 5.1). Under this polarized condition, decoupling methods should perform sub-optimally on the tail class.

5.5 Gleason Grading Data

For our experiments, I will use the PANDA dataset as our open source[1] use case. Aside from the massive number of biopsy samples (∼11,000 slides), I chose this dataset due to it having a 5-class classification task which I believe is sufficient to create long tailed distribution for the purpose of validating our claim. Each biopsy has an accompanying mask that indicates areas of the disease. To keep the experiments tractable, I avoided using the vast majority of the dataset. I will restrict our use to three partitions of the dataset each with 1000 biopsies whose slides are mutually exclusive with other partitions. I divided each partition into a train and test datasets using an 800:200 split, respectively.

Due to the gigapixel size of biopsies, I perform patching operations for the convolutional network to process images. I crop patches in a sliding window manner of size 512x512. Using the accompanying 'Radboud' mask which has the 5-class information, a patch is admitted into the dataset if at least 50% of the area is of a certain class. With the exception of the Stroma class, the patch selection is designed such that no other disease is present other than the disease labeled–avoiding multi-class artifacts.

[1] https://www.kaggle.com/c/prostate-cancer-grade-assessment/data

Table 5.1: Per-class and Average Accuracy (mean±std). Average of three partitions. Bolded figures indicate the largest result when excluding the same class.

Feat. Model	**S**	**B**	**G3**	**G4**	**G5**	**Average**
SvB	**.88+.01**	.61+.04	.30+.01	.25+.07	.36+.07	.48+.04
SvG3	.87+.01	**.50+.07**	.59+.03	**.29+.08**	**.61+.12**	**.57+.06**
SvG4	.83+.02	.49+.04	.32+.08	.34+.05	.32+.07	.46+.05
SvG5	.85+.01	.45+.03	**.47+.05**	.20+.02	.53+.06	.50+.03

5.6 Results

In this section, I summarize the results of our two experiments: In experiment 1, I empirically show that the five-class imbalanced biopsy dataset contains head classes that are sub-optimal to the cancer grading classification task, and, in fact, other classes provide more robust global features to the classification problem; in experiment 2, I demonstrate that the models' performance improves as I proportionally increase the frequency of these key classes.

Datasets and Protocols

I will train models that classify areas of disease using the five classes of the PANDA dataset (S, B, G3, G4 and G5). I perform the experiments and report the average accuracy of the three partitions. For data augmentation, patches are randomly rotated (vertically and horizontally), resized to 112x112 and normalized from 0 to 1. Across the three partitions, I report the results of the per-class and total average accuracy performance of 1000 randomly selected patches in the hold-out test set. For each two experiments, I detail the construction of the training set in their respective sections.

Implementation

For this study, I implement Decouple-cRT[2] (classifier ReTraining) [38]. For the feature model, I use the ResNet-34 architecture. For the classifier, I use a single-layer linear model. The feature model and classifier were trained on 24 and 10 epochs, respectively. In order to avoid introducing feature bias from other datasets, I train every model from scratch using randomly initialized weights. I used a learning rate of $lr = .025$ and decayed by a factor of .1 at every 5 epochs for the feature model, and 2 epochs for the classifier model.

Feature Ablation

In this experiment, I show that the features learned from the default head classes (S and B) are sub-optimal to differentiate the tail classes. I do so by analyzing the effect of the

[2]https://github.com/facebookresearch/classifier-balancing

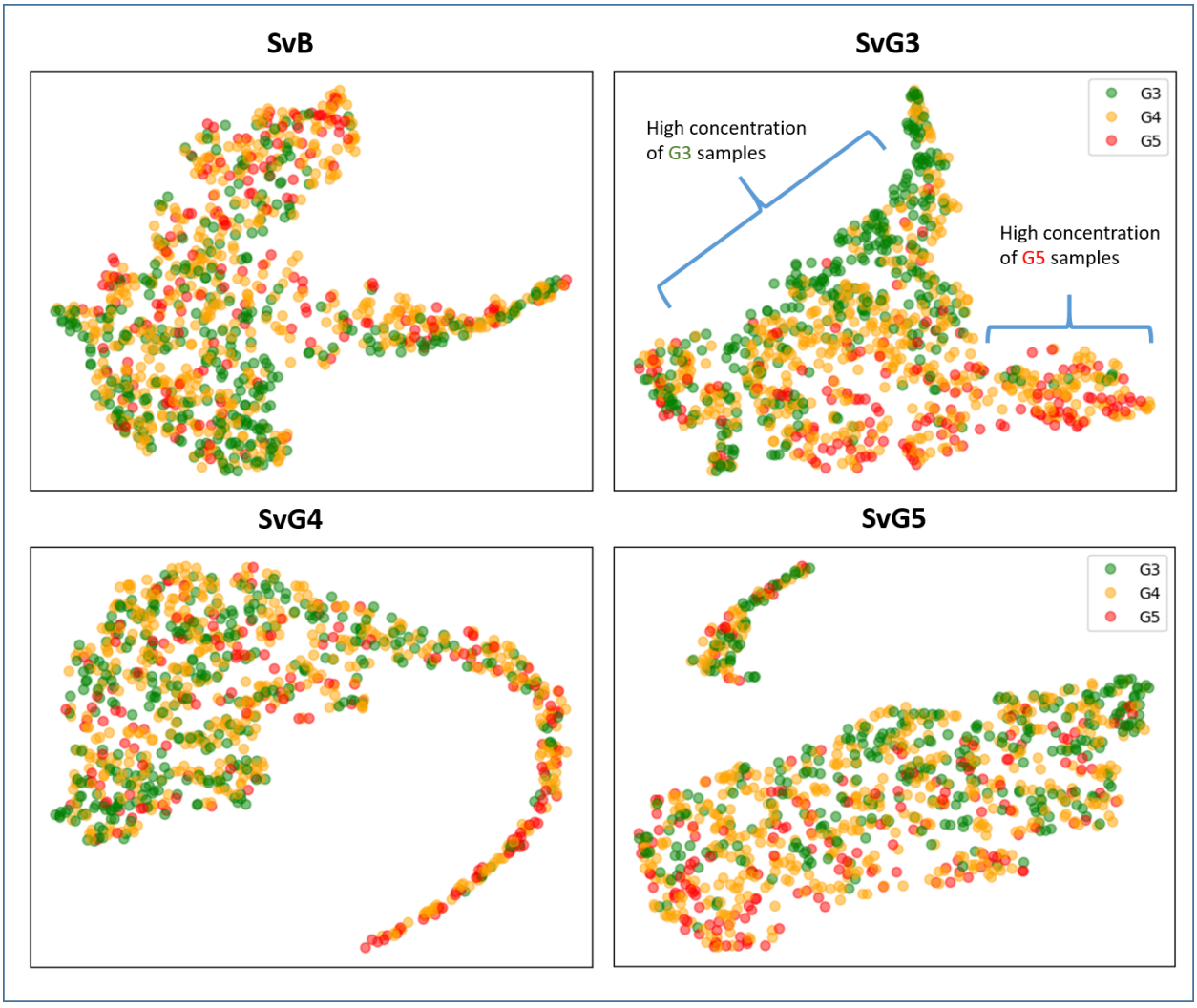

Figure 5.2: t-SNE visualization of the tail G3-G5 features extracted from the feature model. I see that the feature model trained on SvG3 shows stronger ability to discriminate between G3 and G5 classes as G4 samples spans in the space between the two classes. Separation in all other feature models (SvB/G4/G5) is not apparent. Best viewed in color. (perplexity=30, iterations=50,000)

Table 5.2: Class frequency for severe and mild imbalance scenario

λ	**S**	**B**	**G3**	**G4**	**G5**
1.5 (severe)	15,546	3,469	774	173	39
1.25 (mild)	14,298	4,096	1,174	336	96

performance by developing four feature models each trained using only two classes. I alternate training the four terminal classes (B, G3, G4, G5) against the S class. I label these models as SvB, SvG3, SvG4, and SvG5 (i.e. SvB is the feature model trained on S and B classes). This effectively rotates each class to the head. I control for the appearance of other confounding features by not including other classes in the training procedure. During the training phase, I randomly select 2,500 patches for each class to train the feature model and sub-select another 500 patches to train the classifier. The classifier is trained on all 5 classes using the output of the trained feature model as the classifiers input.

Table 5.1 summarizes the result indicating that the SvG3 feature model contains the most distinguishable features in the five-class task as the model results in the highest accuracy by a wide margin (+7%). When excluding results of the class on which each model was trained, I conclude that SvG3 has more robust global features that spans across other classes attaining the highest score for the B, G4, and G5 classes. This indicates that, in the PANDA grading case, G3 contains robust global features that translates well to the tail.

Moreover, Table 5.1 also shows that SvB performs poorly which indicates that the S/B class produces sub-optimal features that cause poor discriminability. This empirically validates our claim that the default head classes do not discriminate well due to polarizing features.

Sensitivity Analysis

In this section, I simulate a long-tailed scenario and observe the change in performance as I proportionally increase the frequency of each class. I experiment with two types of imbalances controlled by parameter $\lambda = \{1.5, 1.25\}$ (which I name 'severe' and 'mild,' respectively) such that the per-class proportion (i.e. the fraction of samples belonging to class x) is:

$$proportion(x) = \frac{exp(-\lambda x)}{\sum_{i \in \mathcal{X}} exp(-\lambda i)} \tag{5.1}$$

where $x \in \mathcal{X} = \{0, 1, 2, 3, 4\}$ is the order of the classes from 0 (Stroma) to 4 (G5). A higher λ accentuates the imbalance, as it increases the sample count of the head class and decreases those from the tail classes. I choose these λ values to be high enough that there are enough samples for G5, and low enough such that the imbalance causes a severe deterioration compared to training using traditional methods. The total size of the training set is 20,000 patches. Table 5.2 summarizes the class count for each scenario.

For this experiment, I proportionally increase the the size of each class and observe the change in performance. I use multipliers instead of fixed increments since multipliers respect

Table 5.3: Per-class and Average Accuracy (mean±std) results for Severe Imbalance ($\lambda = 1.5$). Average of three partitions. Increasing the amount of G3 samples shows the greatest improvement in the the per-class and total performance.

Adjustment	**S**	**B**	**G3**	**G4**	**G5**	**Average**
Baseline-uniform	**.96±.01**	.70±.03	.40±.03	.03±.04	.00±.01	.42±.02
Baseline-crt	.90±.01	.71±.02	.53±.01	.29±.01	.56±.10	.60±.03
x2 S	.91±.01	.69±.01	.50±.02	.26±.07	.53±.01	.58±.02
x2 B	.90±.01	**.75±.01**	.53±.02	.31±.04	**.59±.07**	.62±.03
x2 G3	.90±.01	**.75±.01**	**.59±.01**	**.36±.03**	**.59±.09**	**.64±.03**
x2 G4	.91±.01	.69±.06	.52±.04	.34±.02	.57±.07	.61±.04
x2 G5	.89±.01	.71±.04	.54±.01	.32±.04	.58±.01	.61±.02

the proportional difficulty of attaining certain classes. More specifically, at $\lambda = 1.5$, I make an assumption that doubling G3 patches (774 to 1,548) requires the same amount of effort to that of doubling G4 patches (173 to 346). For our lower bound baselines, we compared the results to Baseline-uniform where were uniformly sampled each class (i.e. no class weighting) and Baseline-crt where we use a two-stage approach.

Table 5.3 summarizes the results showing that increasing G3 improves the classification performance of the model compared to all other change in classes, consistent with the implications in Experiment 1. Moreover, increasing Stroma actually decreases the model performance which indicates further biasing of features learned from the head classes. Also, note that increasing G4 and G5 improves the performance marginally.

Table 5.4 summarizes the mild imbalance results showing consistent outcomes with that of the severe imbalance; although, the margin of improvement decreases for G3. This indicates that the marginal improvement declines as imbalances relax. Finally, on both degrees of imbalance, the Baseline-uniform model performs the poorest.

5.7 Discussion

This study contributes three ideas to the body of literature that helps alleviate the effects of imbalance classes in medical datasets. First, this work confirms that modern 2-stage decoupling methods addressing imbalanced datasets improves classification performance applied to biopsy datasets. Secondly, this work argues against the conventional paradigm that adding more rare cases to the training set will improve model performance. While small improvements do occur, the largest reward comes from focusing efforts on collecting classes with robust global features that may be easier to collect. This work argues that certain classes contain global features that can be transferred to improve the model's ability to predict across all classes. In the PANDA case, G3 contained robust global features that the classifier leverages to discriminate across all classes. Lastly, this work shows that default

Table 5.4: Per-class and Average Accuracy (mean±std) results for Mild Imbalance ($\lambda = 1.25$). Average of three partitions. Increasing the amount of G3 samples shows the greatest improvement in the the per-class and total performance. However, the improvement is marginal as it only increases the performance by 1% over x2 B.

Adjustment	**S**	**B**	**G3**	**G4**	**G5**	**Average**
Baseline-uniform	**.95±.01**	.72±.01	.44±.12	.19±.09	.16±.23	.49±.09
Baseline-crt	.88±.01	.73±.01	.57±.01	.36±.05	.61±.01	.63±.02
x2 S	.89±.01	.73±.01	.59±.01	.36±.04	.60±.01	.64±.01
x2 B	.89±.01	**.77±.01**	.63±.03	.40±.03	.59±.09	.66±.03
x2 G3	.91±.01	.75±.01	**.67±.06**	**.42±.04**	.59±.11	**.67±.04**
x2 G4	.89±.01	.75±.01	.61±.01	.38±.03	**.63±.06**	.65±.02
x2 G5	.88±.03	.73±.03	.57±.01	.35±.05	.61±.04	.63±.03

head classes do not always contain robust features that provide strong differentiability across classes where features are polarized to the extreme end. This is especially true in the Gleason scoring case where cancerous patterns are only present in the tail classes. I believe that this property may also generalize to other diseases diagnosed through the visual examination of biopsies.

5.8 Chapter Summary

This work shows the effectiveness of imbalanced methods on biopsy datasets. I also highlight the polarizing features that can degrade the classification performance as the learned features are biased towards the head classes. I then highlight the idea of classes with robust global features that greatly improve the model performance. Our results show that proportionally increasing the frequency of these classes improves the model performance compared to that when proportionally increasing other classes. While this body of work highlights the existence of global features, it does not provide a way to detect them. In the PANDA case, one hypothesis is that G3 contains robust global features since it is the intermediary class between all the classes and has the greatest probability to contain features that span across all classes. I leave the process of discovering these classes as future work.

Chapter 6

Summary and Conclusions

In this chapter, I conclude the dissertation. In Section 6.1, I begin by reviewing the purpose and goal of this body of work. Next, in Section 6.2, I enumerate the research contributions. In Section 6.3, I describe limitations of this work. Then, finally, in Section 6.4, I touch on future work.

6.1 Review of Purpose and Scope

This body of work aims to explore techniques that minimize the annotation and collection effort for curating datasets. I examine two promising avenues to improve the curation cost of medical imaging datasets: First, using semi-supervised learning methods, I show that it is possible to have a model only trained on a few with a performance that is comparable to that of the fully-supervised case. Second, using two-stage class imbalance methods, I show that it is possible to train biopsy models without the need to collect large expensive class-balanced datasets containing rare data. Because semi-supervised approaches decrease the number of annotation requirements and class imbalance methods decrease the need to collect a balanced amount of data, these methods allow the training of models with less resources necessary than their fully-supervised counterparts.

6.2 Research Contributions

This body of work aims to alleviate the data curation cost of biopsy datasets by addressing the high annotation costs and class imbalances caused by rare diseases common in biopsy applications. This dissertation highlights four scientific contributions to the field of building predictive models with lesser curation effort:

- Contribution 1: I analyze the performance of modern computer vision semi-supervised learning techniques on esophageal [70, 26] and prostate biopsies.

- Contribution 2: I compare and contrast the performance of leading semi-supervised methods on biopsy images. I highlight that current semi-supervised learning research ignores the inherent characteristics of biopsy images–more specifically, the noise contained in biopsy image [70].

- Contribution 3: In order to improve semi-supervised techniques on noisy biopsy images, I evaluate the application of co-teaching [27] within a semi-supervised setting on prostate biopsies. I introduce CoMixMatch which extends the MixMatch [8] semi-supervised learning technique by using a two-model, SSL co-teaching techniques to dampen the negative effects of noisy data. This improves the semi-supervised prediction performance on prostate biopsies.

- Contribution 4: I analyze the effects of two-stage methods on prostate biopsy images for the task of cancer grading. I discover a property unique to biopsy images I term "polarized features" which organizes cancer features at the tail end (i.e. minority classes) of the cancer progression. I show that these polarized features cause sub-optimal performance when applying two-stage methods [38] on imbalanced datasets.

6.3 Limitations

This dissertation does not advocate the total avoidance of labeling or collection. In fact, I achieve sub-optimal results when applying any of the techniques presented. Fully-supervised and balanced methods are still superior. This is important when applied to high-risk situations like medicine. Medical applications still require the need to label and collect as much data as resources will allow. Furthermore, this work does not absolve developers to collect data to build a robust training set to validate these semi-supervised models. However, if there are resource constraints, these methods can produce the best models to minimize the curation effort.

Results of our semi-supervised experiments also show the need to have a diverse amount of patients included in the training set. Given a fixed number of labeled training set, the results show that the performance of semi-supervised methods is maximized by not just increasing the amount of labeled samples, but by also increasing the patient sources.

6.4 Future Work

While this dissertation hopes to decrease the data curation costs of building CNN-based computer-aided diagnostic tools, it only addresses a small portion of the many possible avenues. One way to decrease annotation costs is through the use of pool-based *active learning* [43, 76, 4] which aims to produce a robust model by annotating only the minimum amount of images. Given an unlabeled dataset, active learning uses an iterative process where, in each iteration, the models selects (or queries) a finite set of the unlabeled sample

should a human annotate that would provide the greatest improvement in performance. Indeed, semi-supervised methods has shown to improve active learning querying capabilities [23].

Furthermore, data annotation costs can also be lowered by building better, user-friendly annotation tools. High-grade, well-designed annotation tools that take into account human factors improve the ease of annotation. A common, data-driven way to improve the annotation process is to use pre-trained 'guessing models', trained on previously labeled samples, to *auto-label* unlabeled images. Leveraging the prediction of guessing models, the annotator only needs to correct the prediction. Once a subset is annotated, the corrected images can be used to augment the training data. The main principle is that annotators will not need to start from scratch to label. While the startup cost of labeling the first training samples is high as more errors will occur during the early stages, as more samples are labeled, the easier this labeling process becomes.

To improve this auto-labeling process, one can apply techniques in semi-supervised, active learning, and class imbalance. Semi-supervised learning could improve guessing models by leveraging the use of a larger unlabeled dataset. Active learning can also be used to query the best unlabeled sample that may improve the performance of the next model. And class imbalance methods could improve the auto-labeling of rare diseases.

Future work can also include improvement in representation learning. Researchers have pointed to self-supervised learning in order to improve semi-supervised learning techniques. By first learning a robust representation of the training set, researchers have shown that semi-supervised learning performance can improve. Also, by improving representation, I may also improve the features learned by two-stage methods that address class imbalance.

Bibliography

[1] URL: https://cs231n.github.io/convolutional-networks/.

[2] Dana Angluin and Philip Laird. "Learning from noisy examples". In: *Machine Learning* 2.4 (1988), pp. 343–370.

[3] Guilherme Aresta et al. "Bach: Grand challenge on breast cancer histology images". In: *Medical image analysis* 56 (2019), pp. 122–139.

[4] Jordan T Ash et al. "Deep batch active learning by diverse, uncertain gradient lower bounds". In: *arXiv preprint arXiv:1906.03671* (2019).

[5] Babak Ehteshami Bejnordi et al. "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer". In: *Jama* 318.22 (2017), pp. 2199–2210.

[6] Mikhail Belkin and Partha Niyogi. "Laplacian eigenmaps and spectral techniques for embedding and clustering." In: *Nips*. Vol. 14. 14. 2001, pp. 585–591.

[7] Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux. "Label Propagation and Quadratic Criterion". In: (2006).

[8] David Berthelot et al. "Mixmatch: A holistic approach to semi-supervised learning". In: *Advances in Neural Information Processing Systems*. 2019, pp. 5050–5060.

[9] David Berthelot et al. "ReMixMatch: Semi-Supervised Learning with Distribution Alignment and Augmentation Anchoring". In: *arXiv preprint arXiv:1911.09785* (2019).

[10] Kaidi Cao et al. "Learning imbalanced datasets with label-distribution-aware margin loss". In: *Advances in Neural Information Processing Systems*. 2019, pp. 1565–1576.

[11] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. "Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]". In: *IEEE Transactions on Neural Networks* 20.3 (2009), pp. 542–542.

[12] Nitesh V Chawla et al. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[13] Hanbo Chen et al. "Rectified Cross-Entropy and Upper Transition Loss for Weakly Supervised Whole Slide Image Classifier". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 351–359.

[14] Adam Coates and Andrew Y Ng. "The importance of encoding versus training with sparse coding and vector quantization". In: *ICML*. 2011.

[15] Nicolas Coudray et al. "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning". In: *Nature medicine* 24.10 (2018), pp. 1559–1567.

[16] Pierre Courtiol et al. "Classification and disease localization in histopathology using only global labels: A weakly-supervised approach". In: *arXiv preprint arXiv:1802.02212* (2018).

[17] Ekin D Cubuk et al. "Randaugment: Practical automated data augmentation with a reduced search space". In: *arXiv preprint arXiv:1909.13719* (2019).

[18] Yin Cui et al. "Class-balanced loss based on effective number of samples". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 9268–9277.

[19] Chris Drummond, Robert C Holte, et al. "C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling". In: *Workshop on learning from imbalanced datasets II*. Vol. 11. Citeseer. 2003, pp. 1–8.

[20] Lars Egevad et al. "Standardization of Gleason grading among 337 European pathologists". In: *Histopathology* 62.2 (2013), pp. 247–256.

[21] Andre Esteva et al. "Dermatologist-level classification of skin cancer with deep neural networks". In: *nature* 542.7639 (2017), pp. 115–118.

[22] Alex Gammerman, Volodya Vovk, and Vladimir Vapnik. "Learning by transduction". In: *arXiv preprint arXiv:1301.7375* (2013).

[23] Mingfei Gao et al. "Consistency-based semi-supervised active learning: Towards minimizing labeling cost". In: *European Conference on Computer Vision*. Springer. 2020, pp. 510–526.

[24] Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. "Recent advances in open set recognition: A survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

[25] Arkadiusz Gertych et al. "Convolutional neural networks can accurately distinguish four histologic growth patterns of lung adenocarcinoma in digital slides". In: *Scientific reports* 9.1 (2019), pp. 1–12.

[26] Shan Guleria et al. "Deep learning systems detect dysplasia with human-like accuracy using histopathology and probe-based confocal laser endomicroscopy". In: *Scientific reports* 11.1 (2021), pp. 1–11.

[27] Bo Han et al. "Co-teaching: Robust training of deep neural networks with extremely noisy labels". In: *Advances in neural information processing systems*. 2018, pp. 8527–8537.

[28] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning". In: *International conference on intelligent computing*. Springer. 2005, pp. 878–887.

[29] Haibo He et al. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning". In: *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. IEEE. 2008, pp. 1322–1328.

[30] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[31] Kaiming He et al. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1026–1034.

[32] Kaiming He et al. "Mask r-cnn". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.

[33] Le Hou et al. "Patch-based convolutional neural network for whole slide tissue image classification". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2424–2433.

[34] Chen Huang et al. "Deep imbalanced learning for face recognition and attribute prediction". In: *IEEE transactions on pattern analysis and machine intelligence* 42.11 (2019), pp. 2781–2794.

[35] Peter A Humphrey. "Gleason grading and prognostic factors in carcinoma of the prostate". In: *Modern pathology* 17.3 (2004), pp. 292–306.

[36] Thorsten Joachims et al. "Transductive inference for text classification using support vector machines". In: *Icml*. Vol. 99. 1999, pp. 200–209.

[37] Thorsten Joachims. "Transductive learning via spectral graph partitioning". In: *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*. 2003, pp. 290–297.

[38] Bingyi Kang et al. "Decoupling representation and classifier for long-tailed recognition". In: *arXiv preprint arXiv:1910.09217* (2019).

[39] Sara Hosseinzadeh Kassani et al. "Breast cancer diagnosis with transfer learning and global pooling". In: *arXiv preprint arXiv:1909.11839* (2019).

[40] Sara Hosseinzadeh Kassani et al. "Classification of histopathological biopsy images using ensemble of deep learning networks". In: *arXiv preprint arXiv:1909.11870* (2019).

[41] Salman Khan et al. "Striking the right balance with uncertainty". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 103–112.

[42] Salman H Khan et al. "Cost-sensitive learning of deep feature representations from imbalanced data". In: *IEEE transactions on neural networks and learning systems* 29.8 (2017), pp. 3573–3587.

[43] Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. "Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning". In: *Advances in Neural Information Processing Systems*. 2019, pp. 7024–7035.

[44] Bruno Korbar et al. "Deep learning for classification of colorectal polyps on whole-slide images". In: *Journal of pathology informatics* 8 (2017).

[45] Oren Z Kraus, Jimmy Lei Ba, and Brendan J Frey. "Classifying and segmenting microscopy images with deep multiple instance learning". In: *Bioinformatics* 32.12 (2016), pp. i52–i59.

[46] Alex Krizhevsky, Geoffrey Hinton, et al. "Learning multiple layers of features from tiny images". In: (2009).

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

[48] Miroslav Kubat, Stan Matwin, et al. "Addressing the curse of imbalanced training sets: one-sided selection". In: *Icml*. Vol. 97. Citeseer. 1997, pp. 179–186.

[49] Samuli Laine and Timo Aila. "Temporal ensembling for semi-supervised learning". In: *arXiv preprint arXiv:1610.02242* (2016).

[50] Julia A Lasserre, Christopher M Bishop, and Thomas P Minka. "Principled hybrids of generative and discriminative models". In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. Vol. 1. IEEE. 2006, pp. 87–94.

[51] Dong-Hyun Lee. "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks". In: *Workshop on challenges in representation learning, ICML*. Vol. 3. 2013, p. 2.

[52] Jiahui Li et al. "Signet ring cell detection with a semi-supervised learning framework". In: *International Conference on Information Processing in Medical Imaging*. Springer. 2019, pp. 842–854.

[53] Junnan Li, Richard Socher, and Steven CH Hoi. "Dividemix: Learning with noisy labels as semi-supervised learning". In: *arXiv preprint arXiv:2002.07394* (2020).

[54] Tsung-Yi Lin et al. "Focal loss for dense object detection". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

[55] Tsung-Yi Lin et al. "Microsoft coco: Common objects in context". In: *European conference on computer vision*. Springer. 2014, pp. 740–755.

[56] Geert Litjens et al. "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis". In: *Scientific reports* 6.1 (2016), pp. 1–11.

[57] Jialun Liu et al. "Deep representation learning on long-tailed data: A learnable embedding augmentation perspective". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 2970–2979.

[58] Yun Liu et al. "Detecting cancer metastases on gigapixel pathology images". In: *arXiv preprint arXiv:1703.02442* (2017).

[59] Ziwei Liu et al. "Large-scale long-tailed recognition in an open world". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 2537–2546.

[60] Ming Y Lu et al. "Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding". In: *arXiv preprint arXiv:1910.10825* (2019).

[61] Alexander Selvikvåg Lundervold and Arvid Lundervold. "An overview of deep learning in medical imaging focusing on MRI". In: *Zeitschrift für Medizinische Physik* 29.2 (2019), pp. 102–127.

[62] Xin Luo et al. "Comprehensive computational pathological image analysis predicts lung cancer prognosis". In: *Journal of Thoracic Oncology* 12.3 (2017), pp. 501–509.

[63] Dhruv Mahajan et al. "Exploring the limits of weakly supervised pretraining". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 181–196.

[64] Inderjeet Mani and I Zhang. "kNN approach to unbalanced data distributions: a case study involving information extraction". In: *Proceedings of workshop on learning from imbalanced datasets*. Vol. 126. 2003.

[65] Takeru Miyato et al. "Virtual adversarial training: a regularization method for supervised and semi-supervised learning". In: *IEEE transactions on pattern analysis and machine intelligence* 41.8 (2018), pp. 1979–1993.

[66] Sanjay Mukhopadhyay et al. "Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: a multicenter blinded randomized noninferiority study of 1992 cases (pivotal study)". In: *The American journal of surgical pathology* 42.1 (2018), p. 39.

[67] Yuval Netzer et al. "Reading digits in natural images with unsupervised feature learning". In: (2011).

[68] Yassine El Ouahidi et al. "An Approach for Clustering Subjects According to Similarities in Cell Distributions within Biopsies". In: *arXiv preprint arXiv:2007.00135* (2020).

[69] Mohammad Peikari et al. "A cluster-then-label semi-supervised learning approach for pathology image classification". In: *Scientific reports* 8.1 (2018), pp. 1–13.

[70] J Vince Pulido et al. "Semi-Supervised Classification of Noisy, Gigapixel Histology Images". In: *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE. 2020, pp. 563–568.

[71] Joseph Redmon et al. "You only look once: Unified, real-time object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.

[72] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. "Learning representations by back-propagating errors". In: *nature* 323.6088 (1986), pp. 533–536.

[73] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. "Regularization with stochastic transformations and perturbations for deep semi-supervised learning". In: *arXiv preprint arXiv:1606.04586* (2016).

[74] W. J. Scheirer, L. P. Jain, and T. E. Boult. "Probability Models for Open Set Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.11 (2014), pp. 2317–2324. DOI: 10.1109/TPAMI.2014.2321392.

[75] Walter J Scheirer et al. "Toward open set recognition". In: *IEEE transactions on pattern analysis and machine intelligence* 35.7 (2012), pp. 1757–1772.

[76] Ozan Sener and Silvio Savarese. "Active learning for convolutional neural networks: A core-set approach". In: *arXiv preprint arXiv:1708.00489* (2017).

[77] Nicholas J Shaheen et al. "ACG clinical guideline: diagnosis and management of Barrett's esophagus". In: *Official journal of the American College of Gastroenterology—ACG* 111.1 (2016), pp. 30–50.

[78] Li Shen, Zhouchen Lin, and Qingming Huang. "Relay backpropagation for effective learning of deep convolutional neural networks". In: *European conference on computer vision*. Springer. 2016, pp. 467–482.

[79] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[80] Michael R Smith, Tony Martinez, and Christophe Giraud-Carrier. "An instance level analysis of data complexity". In: *Machine learning* 95.2 (2014), pp. 225–256.

[81] Kihyuk Sohn et al. "Fixmatch: Simplifying semi-supervised learning with consistency and confidence". In: *arXiv preprint arXiv:2001.07685* (2020).

[82] Peter Ström et al. "Pathologist-level grading of prostate biopsies with artificial intelligence". In: *arXiv preprint arXiv:1907.01368* (2019).

[83] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. "Long-tailed classification by keeping the good and removing the bad momentum causal effect". In: *arXiv preprint arXiv:2009.12991* (2020).

[84] Antti Tarvainen and Harri Valpola. "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results". In: *Advances in neural information processing systems*. 2017, pp. 1195–1204.

[85] Ivan Tomek. "Two Modifications of CNN". In: *IEEE Transactions on Systems, Man, and Cybernetics* SMC-6.11 (1976), pp. 769–772.

[86] Naofumi Tomita et al. "Finding a Needle in the Haystack: Attention-Based Classification of High Resolution Microscopy Images". In: *arXiv preprint arXiv:1811.08513* (2018).

[87] Alexander Toshev and Christian Szegedy. "Deeppose: Human pose estimation via deep neural networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 1653–1660.

[88] Jesper E Van Engelen and Holger H Hoos. "A survey on semi-supervised learning". In: *Machine Learning* 109.2 (2020), pp. 373–440.

[89] Oriol Vinyals et al. "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge". In: *IEEE transactions on pattern analysis and machine intelligence* 39.4 (2016), pp. 652–663.

[90] Xi Wang et al. "Weakly supervised learning for whole slide lung cancer image classification". In: *Medical Imaging with Deep Learning* (2018).

[91] Yisen Wang et al. "Iterative learning with open-set noisy labels". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 8688–8696.

[92] Jason W Wei et al. "Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks". In: *Scientific reports* 9.1 (2019), pp. 1–8.

[93] Chiyuan Zhang et al. "Understanding deep learning requires rethinking generalization". In: *arXiv preprint arXiv:1611.03530* (2016).

[94] Hongyi Zhang et al. "mixup: Beyond empirical risk minimization". In: *arXiv preprint arXiv:1710.09412* (2017).

[95] Zizhao Zhang et al. "Pathologist-level interpretable whole-slide cancer diagnosis with deep learning". In: *Nature Machine Intelligence* 1.5 (2019), pp. 236–245.

[96] Boyan Zhou et al. "Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 9719–9728.

[97] Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. "Semi-supervised learning using gaussian fields and harmonic functions". In: *Proceedings of the 20th International conference on Machine learning (ICML-03)*. 2003, pp. 912–919.