Image Segmentation in Histopathology with Limited Labeled Data

---

A

Dissertation

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

---

in partial fulfillment

of the requirements for the degree

Doctor of Philosophy

by

Payden McBee

August 2023

APPROVAL SHEET

This

Dissertation

is submitted in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

Author: Payden McBee

This dissertatin has been read and approved by the exdaming committee

Advisor: Donald Brown

Advisor:

Committee Member: Laura Barnes

Committee Member: Michael Porter

Committee Member: Sana Syed

Committee Member: William Adorno

Committee Member:

Committee Member:

Accepted for the School of Engineering and Applied Science:

Jennifer L. West, School of Engineering and Applied Science

August 2023

Image Segmentation in Histopathology with Limited Labeled Data

Abstract
Image Segmentation in Histopathology with Limited Labeled Data

by

Payden McBee

Doctor of Philosophy in Systems and Information Engineering

University of Virginia

Detecting and quantifying cellular features via deep neural networks informs understanding of disease progression in digital histopathology. Biopsies from patients are placed on slides, then stained and digitized to create whole-slide images (WSIs). Medical experts provide pixel-level and image-level annotations that outline cell types and disease information on image patches from the WSI. Segmentation neural networks trained on these annotations provide pixel-wise predictions of cell types and tissue disease. However, due to the large amount of tissue imaged for every biopsy, medical experts only label a small portion of the imagery per patient. This results in a large amount of unlabeled data that standard supervised algorithms cannot use. The main question of this research is: How do we optimize segmentation performance in histopathology in limited labeled data settings?

First, model initialization in limited data settings is addressed. Transfer learning techniques have proven more beneficial than random initialization of model weights in many settings. In transfer learning, a model developed for a specific task is reused as the initial model for a second task with limited labeled data. Many models are pre-trained on natural image sets, such as ImageNet, and fine-tuned on medical images. Additionally, some models are pre-trained on one set of medical images and fine-tuned on another set. However, most models are only pre-trained with ImageNet, and there is no standardized medical equivalent of ImageNet. Thus, what type of model initialization is optimal in limited data settings? My published results show the optimality of model weights pre-trained with ImageNet over those pre-trained with histopathology images when the labeled dataset is small. This allows for a broader range of architectures, saving time and preventing expensive gathering of histopathology data and pre-training.

Second, I consider how unlabeled data can be used to optimize segmentation performance. Pseudo-labeling is an existing semi-supervised learning technique that uses unlabeled data to train a model. Existing techniques utilize confidence and uncertainty quantification to select images for pseudo-labeling from a classification standpoint, but the literature does not extend them to the segmentation context. Furthermore, techniques that use pseudo-labels for segmentation either do not specify how the unlabeled data was selected or use a deterministic threshold. Due to the large amount of unlabeled imagery, using all of the WSI in training is not practical. The literature does not address the trustworthiness of the model's uncertainty quantification. Thus, the second contribution of this research is to adapt and verify the utility of confidence and uncertainty quantification methods from a classification setting to the segmentation setting and to inform image-level

selection. My published results show the importance of assessing the correlation between the image-level uncertainty metric and the model performance on a labeled set as a precondition for using the model to select unlabeled images. My approach enables the prioritization of images that maximize performance and provide trust in the model via an intuitive visualization of uncertainty.

Third, I consider how biases inherent in unlabeled and labeled data can be identified that would hinder the generalization of semi-supervised algorithms. Existing methods address differences in labeled and unlabeled sets but do not provide clinically actionable interpretations. I use Gaussian Mixture Models to cluster the unlabeled and labeled sets to identify sampling and labeling biases and demonstrate the effect of these biases in semi-supervised learning algorithms. My method provides clear interpretability about biases that enables the correct clinical solution, reducing cost and minimizing procedures necessary for the patient.

A Marina, mi querida

Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Image Classification Levels

In computer vision, there are different levels of image classification. Figure 1.1 shows four types: image classification, object localization, instance segmentation, and semantic segmentation. In the context of cell detection, the simplest level of an image classification algorithm outputs a probability determining whether the image contains cells or not. Object localization predicts a bounding box containing each object or cell of interest. Instance segmentation identifies the pixels of each distinct cell, such as the pixels of cell 1, cell 2, et cetera. The focus of this research is the last type, semantic segmentation. Each pixel is assigned a probability of belonging to a given cell class or the background class. To do this, a segmentation model, such as a U-Net (Ronneberger et al., 2015), trains on a set of images and the corresponding pixel-wise class annotations and then, given a new image, predicts the class predictions per pixel.

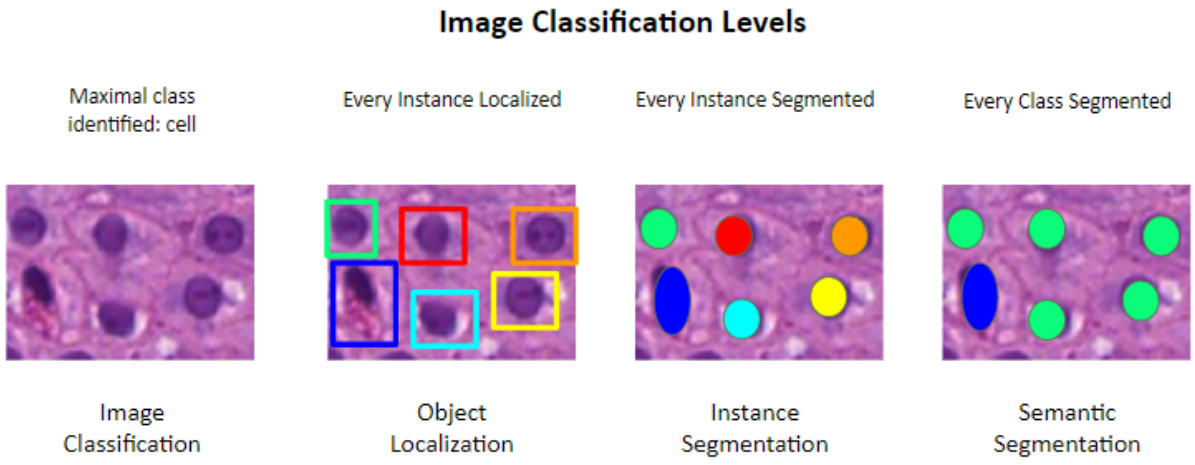

Figure 1.1 Types of Image Classification: This figure shows 4 image classification levels and their associated annotations.

1.2 Digital Histopathology

In addition to the engineering focus of semantic segmentation within computer vision, the medical background of this research includes digital histopathology. Digital histopathology concerns the study of digitized pictures of tissue with the intention of better understanding underlying diseases. Biopsies are taken from a patient to create these

digitized images.

Tissue samples are then placed on slides, stained with hematoxylin and eosin, and imaged to create a whole-slide-image (WSI). WSIs are large, containing multiple tissue regions where each region is on the order of 10,000 x 10,000 pixels. Medical professionals then provide pixel-wise annotations for a small portion of the imagery to train a model. This results in a large amount of the WSI that is left unlabeled. Given the labeled data, a model can be trained and used to quantify cell count and prevalence, which informs understanding of disease progression. Figure 1.2 shows an image patch from a patient with esophageal esophagitis, where medical professionals have annotated the eosinophils and a model's output.

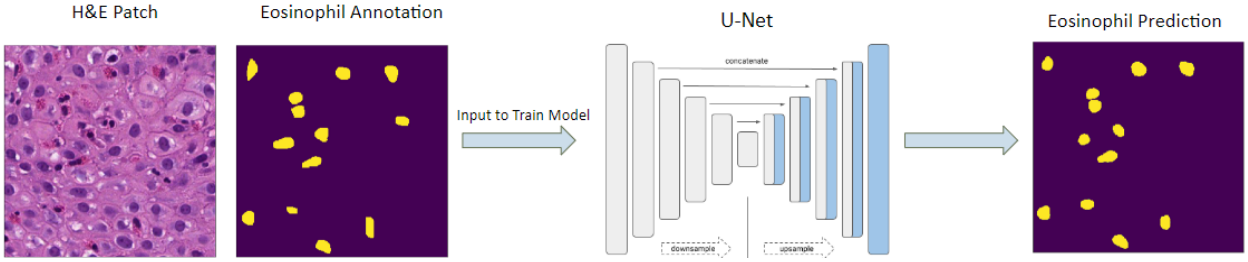

Figure 1.2 Histopathological Segmentation Example: This figure shows a U-Net being trained on a 3-channel RGB image of tissue stained with hematoxylin and eosin sampled from the esophagus of a patient with eosinophilic esophagitis with annotations for eosinophils. The output of the model is the segmentation map for the predicted locations of the eosinophils.

1.3 The Problem

Next, I define the problem this research is addressing. Since segmentation labeling requires a significant time commitment from experts, not all tissue can be labeled. This leaves large portions of unlabeled tissue that traditional supervised learning algorithms cannot utilize in training. Furthermore, no large segmentation labeled histopathology dataset exists on the scale that does for common objects. Thus, the main research question is: How do we maximize segmentation performance in histopathology in limited labeled data settings? This work breaks this problem down into subcomponents: model initialization, incorporation and trustworthiness of unlabeled data, and identification and evaluation of bias in unlabeled and labeled datasets. When training a segmentation model, one of the first issues to address is how to initialize the model weights. Current techniques leverage transfer learning, where a model is trained on a large labeled dataset and then fine-tuned on the smaller target dataset. Since no such large labeled dataset exists in histopathology, multiple smaller histopathological segmentation datasets are examined in the transfer learning context, as well as a large, mostly unlabeled histopathological dataset. Second, I investigate using the unlabeled data to increase the segmentation model performance. The literature shows that pseudo-labeling can leverage unlabeled data to increase model performance. However, in the context of whole-slide-imagery, using the entire unlabeled set would be infeasible due to its immense size. Thus, a method to select a subset of images from the unlabeled set for pseudo-labeling is needed. Existing techniques use empirical thresholds or expensive pixel-level uncertainty quantification. Current

techniques also do not address the trustworthiness of the model's uncertainty metric on unlabeled data. Third, differences between the labeled and unlabeled data can degrade the performance of segmentation algorithms. Existing techniques identify these differences and even correlate the size of the difference with performance. However, there is no root cause analysis that identifies the biases that occur in the data that lead to the differences in distributions.

1.4 Dissertation Overview

The following is an overview of the findings and impact of this research in answering the main problem, broken out by the subcomponents.

1. How should models be initialized for histopathological segmentation in limited labeled data settings?

   Contribution: Showed optimality of ImageNet pre-trained weights over histopathology weights when the dataset is small.

   Impact: Allows for a wider range of architectures, saving time and preventing expensive gathering of histopathology data and pre-training.

2. How can unlabeled data be incorporated and trusted for image-level selection?

   Contribution: Adapted confidence and uncertainty quantification methods from classification to segmentation setting for image-level selection and leveraged uncertainty correlation on the training set.

   Impact: Enables prioritization of images that maximize performance and provide trust in the model via intuitive uncertainty visualization.

3. How can biases inherent in unlabeled and labeled data be identified that would hinder the generalization of semi-supervised algorithms?

   Contribution: Used clustering to identify sampling and labeling biases and demonstrated effect on semi-supervised learning techniques.

   Impact: Gives clear interpretability about biases that enables the correct clinical solution, reducing cost and minimizing unnecessary procedures for patients.

Chapter 2
Segmentation Datasets

2.1 Histopathological Datasets

In histopathological image analysis, biopsies are typically stained with hematoxylin and eosin (H&E). The images analyzed in this work, as is the case with most publicly available histopathological datasets, are H&E stained (Komura and Ishikawa, 2018). Hematoxylin stains the nuclei in cells blue, and eosin stains the cytoplasm and extracellular matrix features varying shades of pink (Fischer et al., 2008). For the purposes of segmentation, medical professionals provide labels for cell types and cellular features for an algorithm to learn. In this work, 5 different histopathological segmentation datasets are examined, as described in the following sections.

2.1.1 Eosinophilic Esophagitis (EoE) Dataset

The EoE dataset consists of images from 30 patients diagnosed with eosinophilic esophagitis (EoE), where the biopsies are taken from the esophagus during an endoscopy. EoE is a progressive disease that presents as episodes of vomiting, dysphagia, and heartburn. Chronic inflammation and tissue remodeling can lead to luminal narrowing due to the formation of strictures and extensive fibrosis. The gold standard for diagnosis is tissue biopsy with greater than or equal to 15 intraepithelial eosinophils per high power field on light microscopy (Aceves 2011; Gonsalves and Aceves, 2020). Eosinophils are proinflammatory cells with bilobed nuclei which appear as two nuclei pressed into one another to make a figure 8 pattern. The cytoplasm is filled with red-staining secretory granules on H&E staining (Rosenberg et al., 2013). Tissue eosinophils can be seen on light microscopy in various states of degranulation. Intact eosinophils are annotated in this dataset. The EoE dataset consists of 514 labeled and 240,526 unlabeled images, each with a size of 512 x 512 pixels at 40x magnification. The use of the EoE data was approved under IRB-HSR 19562, Eosinophilic Esophagitis Patient Database and Biorepository.

2.1.2 Crohn's Disease (CD) Dataset

The CD dataset consists of images from 51 patients, where the biopsies were taken during a colonoscopy, with varying levels of disease presence. Each image is annotated for eosinophils. Eosinophil density may correlate with disease activity in other inflammatory gastrointestinal diseases, such as Crohn's disease and ulcerative colitis (UC) (Alhmoud et al., 2020). The CD dataset consists of 200 labeled images and 291,779 unlabeled images, each with a size of 512 x 512 pixels at 40x magnification.

2.1.3 Colorectal Nuclear Segmentation and the Phenotypes (CoNSeP) Dataset

The CoNSeP dataset consists of images from 16 patients suffering from colorectal adenocarcinoma (CRA) (Graham et al., 2019). Each image is annotated for 6 classes of nuclei: normal epithelial, tumor epithelial, inflammatory, necrotic, muscle, and fibroblast. The CoNSeP dataset consists of 41 labeled images of size 1000 x 1000 pixels at 40x magnification.

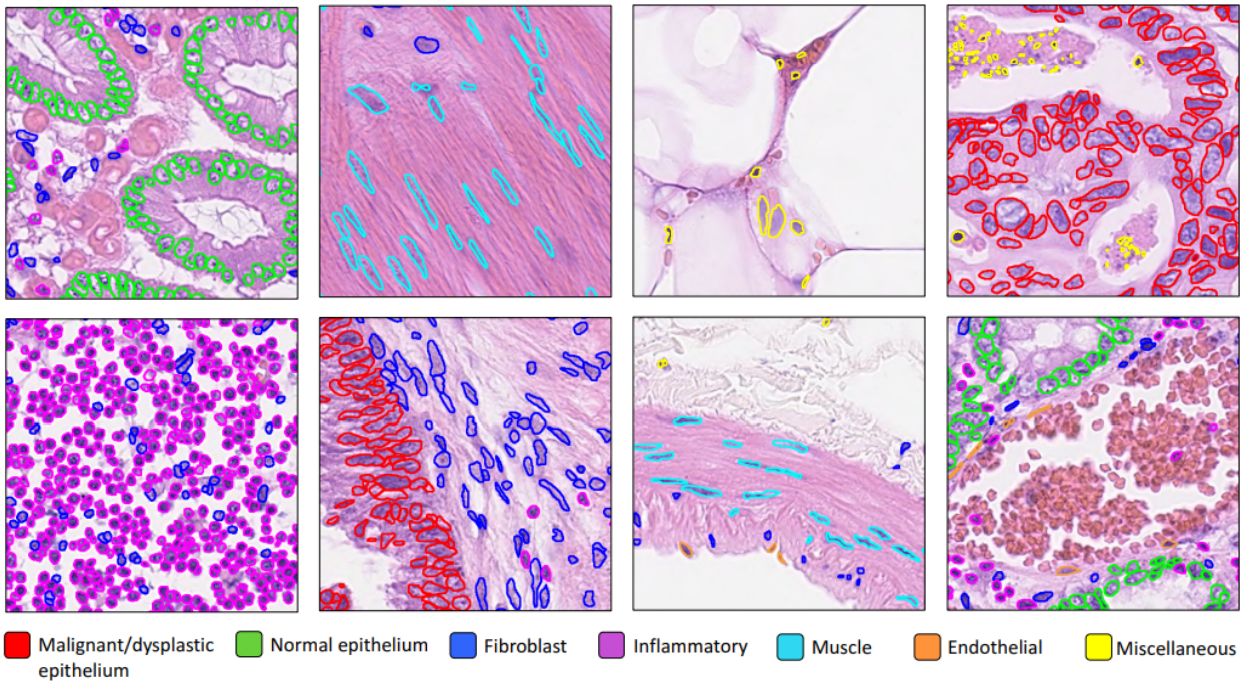

Figure 2.1: Patches from CoNSeP dataset from Graham et al. (2019)

2.1.4 PanCancer Histology Dataset for Nuclei Instance Segmentation and Classification (PanNuke)

PanNuke consists of images from 19 different organs with various kinds of cancer. (Gamper et al., 2019). Each image is annotated for 5 classes of nuclei: neoplastic, inflammatory, connective, dead, and non-neoplastic epithelial. PanNuke consists of 1,485 labeled images of size 256 x 256 pixels at varying levels of magnification.

2.1.5 Multi-Organ Nuclei Segmentation (MoNuSeg) Dataset

The MoNuSeg dataset consists of images from 30 patients and 7 different organs, such as the kidney and liver (Kumar et al., 2020). Each image was derived from one WSI per patient from The Cancer Genomic Atlas (TCGA 2016). The dataset is annotated for nuclei and non-nuclei pixels. The MoNuSeg dataset consists of 44 images of size 1000 x 1000 pixels at 40x magnification. The pre-processing of each dataset is described in the following chapters. Of note, each of the labeled datasets are relatively small, with PanNuke having the largest labeled set of only 1,495 images. In the following chapter, how to address model initialization with small labeled datasets is addressed.

Chapter 3
Model Initialization

3.1 Summary

In limited data settings, transfer learning has proven useful in initializing model parameters. In this chapter, I compare random initialization, pre-training on ImageNet, and pre-training on histopathology datasets for 2 model architectures across 4 segmentation histopathology datasets. I show that pre-training on histopathology datasets does not always significantly improve performance relative to ImageNet pre-trained weights for both model architectures. I conclude that unless larger labeled datasets or semi-supervised techniques are leveraged, ImageNet pre-trained weights should be used in initializing segmentation models for histopathology.

3.2 Literature Review

Transfer learning is a technique where a model developed for a specific task can be reused as the initial model for a second task with limited labeled data. A common transfer learning approach for medical images is to start with the standard network architectures, e.g., VGG (Simonyan and Zisserman, 2014) and ResNet (He et al., 2016) pre-trained on the large-scale natural images such as ImageNet (Deng et al., 2009) and PASCAL VOC (Everingham et al., 2010), and then fine-tune them on medical images, as in Figure 3.1.

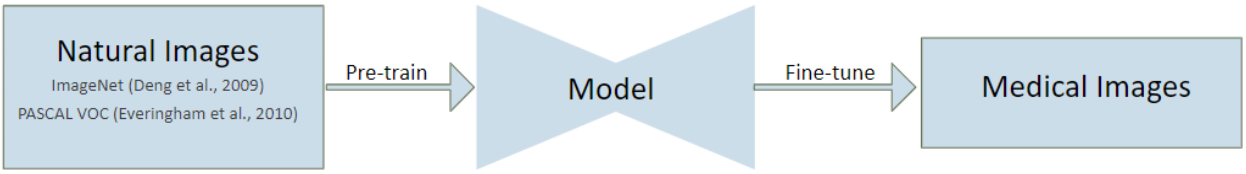

Figure 3.1: This figure shows the transfer learning process of a model training on natural image datasets and being fine-tuned on medical images.

The effectiveness of pre-trained deep convolutional neural networks (CNNs) with sufficient fine-tuning was investigated on four medical imaging applications in Tajbakhsh et al. (2016). This study demonstrated that, in most cases, fine-tuning a pre-trained model achieved better performance and robustness than those trained from scratch with random initialization. Similarly, Devan et al. (2019) demonstrated that transfer learning with ImageNet can significantly enhance model performance in detecting herpesvirus capsids in microscopy images, particularly when labeled data is limited. Conze et al. (2020) utilized a VGG-11 encoder pre-trained on ImageNet for the shoulder muscle MRI segmentation task. These results indicate that a CNN pre-trained on ImageNet learns features that are

applicable to both natural and medical images. However, the gap in features between medical and natural images has motivated pre-training on medical datasets. Ray et al. (2022) demonstrated an increase in performance and faster convergence for CNNs pre-trained on histopathological datasets relative to a model pre-trained on ImageNet. Similarly, Ciga et al. (2022) used self-training on unlabeled histopathology images to improve segmentation results.

3.3 Discussion and Limitations of Literature

Ray et al. (2022) and Ciga et al. (2022) demonstrate pre-training on histopathological datasets as superior to pre-training on ImageNet in classification and segmentation, respectively. In each case, large histopathological datasets are gathered for pre-training and the model weights are trained. Specifically, Ray et al. (2022) pre-trained with over 100,000 images distinguishing patches of cancer tissue from normal tissue. Ciga et al. (2022) pre-trained with 400,000 images from several sources, mostly unlabeled.

This gathering of histopathological data was required because no standardized version exists, nor do they approach the size of ImageNet, with 14 million labeled images. Additionally, most models are not pre-trained on histopathological data, though they are usually pre-trained on ImageNet. Thus, to use a new architecture, not only does a histopathological dataset need to be assembled, but the architecture itself has to be trained on it. Finally, pre-training on labeled histopathological data has not been tested for segmentation, only encoders for classification tasks or via unlabeled self-supervision as in Ciga et al. have been explored. Specifically, the efficacy of utilizing a model pre-trained on natural images compared to a medical image pre-trained model for nuclei segmentation tasks has not been investigated.

In light of these limitations, this section seeks to answer the following: Does pre-training on histopathology datasets improve segmentation performance relative to encoders pre-trained on ImageNet?

The contribution of this chapter answers this question by showing the optimality of ImageNet pre-trained weights over histopathology weights when the dataset is small. The impact of this contribution allows for a wider range of architectures to be leveraged, saving time and preventing the need for expensive gathering of histopathology data and pre-training.

3.4 Pre-processing of Datasets

First, I discuss the pre-processing of 4 datasets derived from whole-slide-images in my evaluation. The eosinophilic esophagitis (EoE) labeled dataset consists of 514 images at 512 x 512 pixels. Each image was divided into 4 via a sliding window with no overlap and resized to 256, creating 2,056 images, which are annotated for eosinophils. The 200 images in the Crohn's Disease (CD) dataset are processed similarly to produce 800 images. The Colorectal Nuclear Segmentation and the Pheno-types (CoNSeP) dataset (Graham et al., 2019) was cropped to create 660 images. The PanCancer Histology Dataset for Nuclei Instance Segmentation and Classification (PanNuke) (Gamper et al., 2019) was cropped into 7,901 images.

3.5 Methodology

For this analysis, the HoVer-Net (Graham et al., 2019) and U-Net++ (Zhou et al., 2018) models are used. HoVer-Net has three separate task-specific decoders, which are used for nuclei detection, separation, and classification, respectively. U-Net++ has a single decoder to provide pixel classification. The Preact-ResNet50 is utilized as the encoder for the HoVer-Net model and ResNet50 as the encoder for the U-Net++ model. For HoVer-Net, the hyper-parameters and training strategies presented by Graham et al. (2019) are followed. For U-Net++, each model is trained for 200 epochs and select the model that minimizes the validation binary cross-entropy (EoE and Crohns) or cross-entropy (CoNSeP and PanNuke) loss. Each of the models is trained and tested for each of the 4 datasets given encoders with various pre-trained weights. The MoNuSAC ResNet50 encoder weights from (Graham et al., 2019) are used, and the other weights are obtained by initializing a model with ImageNet and training it on a given histopathology dataset.

3.6 Experiments and Results

Table 3.1 shows the average performance of the U-Net++ and HoverNet models over 3 runs across the various pre-trained weights for EoE, Crohns, PanNuke, and CoNSeP. I put the maximum performance for each test set and model across the pre-trained weights in bold and put a star if optimal performance is statistically significant. Notably, the models pre-trained on histopathology and the models pre-trained on ImageNet do not have differences that are statistically significant, except for HoVer-Net pre-trained on MoNuSAC for PanNuke, where $p = 0.052499$ from a Welch's t-test comparing it with the ImageNet pre-trained model performance. This indicates that pre-training on these histopathology datasets does not increase the segmentation performance relative to ImageNet weights. The randomly initialized weights are lower for all datasets except U-Net++ for EoE, suggesting that some kind of pre-training is useful. Furthermore, the number of epochs trained when using a model initialized with ImageNet weights is comparable to models pre-trained on histopathology, being significantly larger only for U-Net++ on PanNuke. Thus, there is no set of consistently optimal pre-trained weights, and the ImageNet weights provide the same or better performance than weights from a model pre-trained on multiple histopathology datasets. Also, the time for training for models with ImageNet pre-trained encoder is comparable to models pre-trained with histopathology.

3.7 Conclusion

Training a model with ImageNet pre-trained weights was shown to not have a significantly different performance than pre-training on multiple histopathology datasets for 2 state-of-the-art medical segmentation models, the U-Net++ and HoVer-Net. This is likely partly due to the relatively small size of the datasets used in pre-training. Small datasets do not allow the model to learn diverse features, even when they come from the target domain. Furthermore, the number of training epochs to minimize the validation loss did not increase for the models pre-trained with ImageNet relative to those trained on

histopathology. Unless an abundant amount of histopathology data is available, pre-training on relatively small histopathology datasets is not likely to increase performance or decrease training time relative to an ImageNet baseline.

Table 3.1 Pre-training Model Performance: This table shows the average performance of the U-Net++ and HoverNet models over 3 runs across the various pre-trained weights for EoE, Crohns, PanNuke, and CoNSeP.

		Test Dataset			
	Pre-Trained	Crohns		EoE	
Model	Weights	Dice	Epochs	Dice	Epochs
U-Net++	Random	0.55 ± 0.033	84	**0.62** ± 0.009	83
U-Net++	ImageNet	**0.572** ± 0.006	31	0.615 ± 0.02	60
U-Net++	MoNuSAC	0.554 ± 0.009	109	0.612 ± 0.015	103
U-Net++	CoNSeP	0.565 ± 0.019	30	0.618 ± 0.018	63
U-Net++	PanNuke	0.554 ± 0.031	11	0.599 ± 0.001	65
U-Net++	EoE	0.557 ± 0.026	21	-	-
U-Net++	Crohns	-	-	0.606 ± 0.016	89
HoVer-Net	Random	0.389 ± 0.192	74	0.572 ± 0.007	80
HoVer-Net	ImageNet	**0.609** ± 0.012	90	0.621 ± 0.004	93
HoVer-Net	MoNuSAC	0.6 ± 0.011	83	**0.624** ± 0.002	97
		PanNuke		CoNSeP	
U-Net++	Random	0.552 ± 0.014	73	0.65 ± 0.011	129
U-Net++	ImageNet	0.571 ± 0.013	127	0.669 ± 0.008	82
U-Net++	MoNuSAC	0.54 ± 0.012	69	0.667 ± 0.009	109
U-Net++	CoNSeP	0.57 ± 0.02	50	-	-
U-Net++	PanNuke	-	-	**0.678** ± 0.012	62
U-Net++	EoE	**0.593** ± 0.033	114	0.656 ± 0.039	83
U-Net++	Crohns	0.577 ± 0.017	92	0.664 ± 0.022	81
HoVer-Net	Random	0.555 ± 0.007	78	0.403 ± 0.048	67
HoVer-Net	ImageNet	0.585 ± 0.003	94	0.67 ± 0.007	97
HoVer-Net	MoNuSAC	**0.602*** ± 0.008	98	**0.679** ± 0.016	83
HoVer-Net	CoNSeP	0.589 ± 0.003	95	-	-
HoVer-Net	PanNuke	-	-	0.644 ± 0.05	78

Chapter 4

Incorporating Unlabeled Data and Verifying Trustworthiness

4.1 Summary

Advancements in deep learning techniques have proved useful in biomedical image segmentation. However, the large amount of unlabeled data inherent in biomedical imagery, particularly in digital pathology, creates a semi-supervised learning paradigm. Specifically, because of the time-consuming nature of producing pixel-wise annotations and the high cost of having a pathologist dedicate time to labeling, there is a large amount of unlabeled data that can be used to train segmentation algorithms. Pseudo-labeling is one method to leverage the unlabeled data to increase overall model performance. In this chapter, I adapt a method used for image classification pseudo-labeling to select images for segmentation pseudo-labeling and apply it to 3 digital pathology datasets. To select images for pseudo-labeling, I create and explore different thresholds for confidence and uncertainty on an image-level basis. Furthermore, I study the relationship between image-level uncertainty and confidence with model performance. I find that the certainty metrics do not consistently correlate with performance intuitively, and abnormal correlations serve as an indicator of a model's ability to produce pseudo-labels that are useful in training.

4.2 Literature Review

After the model has been initialized, the next question is how to incorporate unlabeled images. All image patches are extracted from large whole-slide images (WSIs). Each WSI is on the order of multiple gigabytes of data, and the segmentation annotations are inherently cumbersome. Thus, a pathologist is only able to annotate a small portion of a whole slide image per patient, leaving most of the WSIs unlabeled and unused. This context invites the use of semi-supervised learning (SSL), which seeks to leverage both labeled and unlabeled data to increase a model's performance on a hold-out test set. One tool to utilize this unlabeled data is pseudo-labeling.

4.2.1 Pseudo-Labels

Pseudo-labels were introduced by Lee (2013). Figure 4.1 shows the general flow of using pseudo-labels. First, a model is trained on a labeled set. Then, the model is used to predict the labels for the unlabeled data. The class, $i \in C$, with the maximum probability of the model's predictions over all classes is the pseudo-label for a given input.

Equation 1 from Lee (2013) represents this as a one-hot embedding of each class, *i*.

$$y_i' = \begin{cases} 1 & \text{if } i = \text{argmax}_{i'}\, f_{i'}(x) \\ 0 & \text{otherwise} \end{cases} \qquad \text{(Equation 1)}$$

These pseudo-labels are then used along with the labeled set to train a new model via the joint loss function in Equation 2, where n is the number of samples and C is the number of classes.

$$L = \frac{1}{n}\sum_{m=1}^{n}\sum_{i=1}^{C} L(y_i^m, f_i^m) + \alpha(t)\frac{1}{n'}\sum_{m=1}^{n'}\sum_{i=1}^{C} L(y_i'^m, f_i'^m) \qquad \text{(Equation 2)}$$

Lee (2013) used an annealing process to linearly increase the weight of the pseudo-labels in the loss function over time, increasing the influence of the pseudo-labels as the model grew more confident. Equation 3 describes this linear weighting by a stepwise function $\alpha(t)$.

$$\alpha(t) = \begin{cases} 0 & t < T_1 \\ \frac{t-T_1}{T_2-T_1}\alpha_f & T_1 \le t < T_2 \\ \alpha_f & T_2 \le t \end{cases} \qquad \text{(Equation 3)}$$

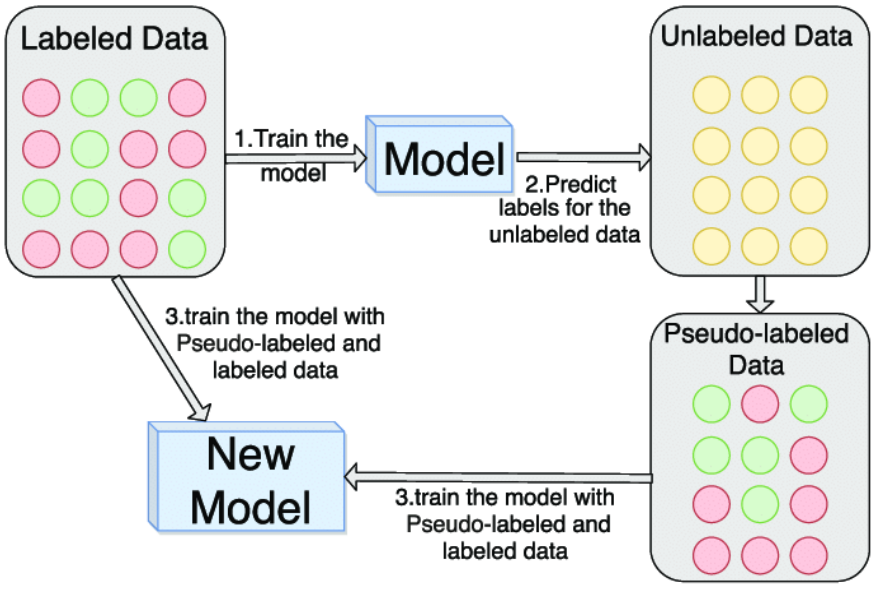

Figure 4.1: This figure shows the process of pseudo-label training from AlZoubi et al. (2020).

4.2.2 Theory of Pseudo-Labels

Lee (2013) describes the rationale behind pseudo-labeling as entropy regularization, where the conditional entropy of class probabilities for the unlabeled data is minimized. The conditional entropy of the unlabeled data measures the amount of class overlap. Thus, minimizing the conditional entropy of the class probabilities minimizes the class overlap of the unlabeled data in the embedding space. This minimization results in a low-density separation between the classes. Figure 4.2 from Lee (2013) shows the t-SNE 2D embedding of the network output on a test set that was trained with (a) and without (b) pseudo-labels. The plot on the right shows how the entropy is minimized by the clear separation between classes, achieved when training with pseudo-labels.

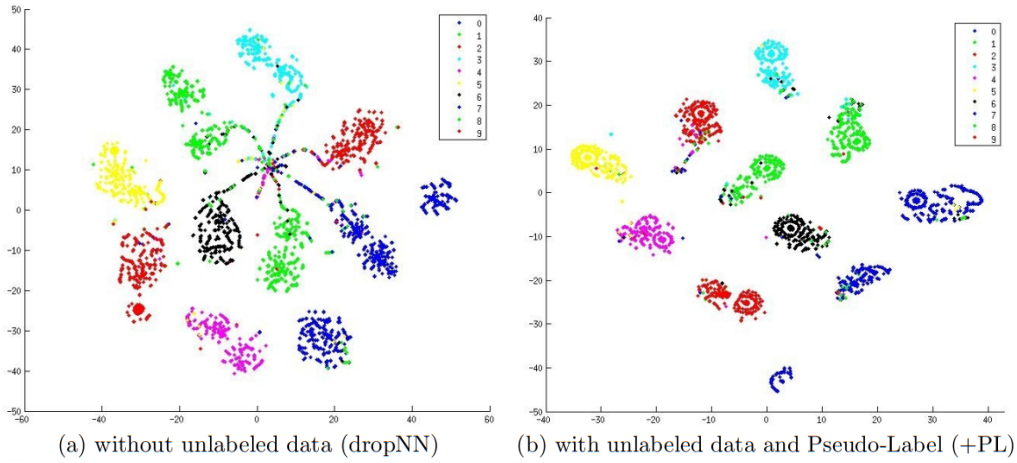

Figure 4.2: This figure shows the embedding of the network's output given test data from Lee (2013).

4.2.3 Selecting and Using Pseudo-Labels in Image Classification

Now, I examine how techniques in the literature select images to pseudo-label and use them in training. From Equation 3, Lee (2013) slowly increases the weight of the pseudo-labels in the loss over time, but does define which data points to choose. Lee (2013) assumes that all unlabeled data should be used.

From the image classification literature, techniques exist to help select images in pseudo-labeling and incorporating them into a semi-supervised training process. I provide a brief overview of each technique and then thoroughly explain each one. The first technique is deterministic confidence thresholding, as defined in the FixMatch algorithm (Sohn et al., 2020), which only selects images if the model prediction has a probability greater than a given threshold, found empirically. The next technique, called FlexMatch (Zhang et al., 2021), builds upon FixMatch by lowering the confidence threshold for less common classes, inversely proportional to the class frequency in the pseudo-labeled set. This increases the likelihood of selecting less common classes for pseudo-labeling, resulting in a model that learns from a more balanced distribution. The last classification technique, from Rizve et al. (2020), selects images for pseudo-labeling that meet both confidence and

uncertainty thresholds. The model produces an uncertainty score for each image, which must be greater than an empirical threshold to be selected. Finally, I consider Zheng et al. (2021), which uses adversarial examples to create a diverse ensemble of models. It also uses pixel-wise uncertainty weighting in the loss functions for both labeled and unlabeled data.

FixMatch: Simplifying semi-supervised learning with consistency and confidence (Sohn et al., 2020)

In FixMatch, pseudo-labels are used in a consistency regularization framework. An image is weakly augmented (only flipped and shifted), and a trained model predicts the class probabilities. If the maximum probable class is greater than a deterministic threshold, the pseudo-label is created as the one-hot encoding of the maximum probability class. That same image is strongly augmented via heavy distortions such as color distortions and shearing. The cross-entropy loss of the prediction probabilities and the pseudo-label is computed. This loss, along with the cross entropy loss of the weakly augmented labeled data are used to train the model. Figure 4.3 shows a diagram of this process.

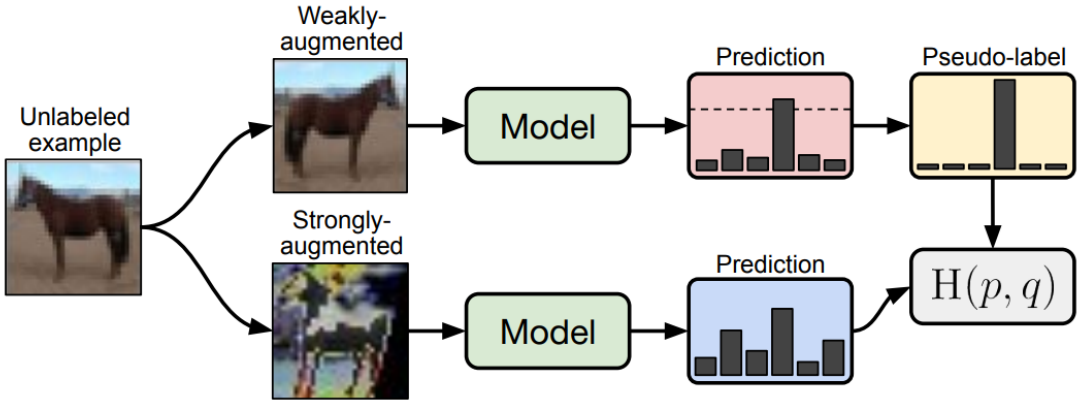

Figure 4.3: This figure shows the training process in the FixMatch algorithm using unlabeled training data from Sohn et al. (2020).

One issue with translating this method to the segmentation context is that shearing, the strong augmentation used, cannot be applied. This is because physical distortions of the pixel-space would change the class of the pixels, and the wrong pixels would be matched in the entropy minimization. Thus, the augmentation techniques would be limited to non-geometric ones, such as color shifts.

FlexMatch: Boosting Semi-Supervised Learning with Curriculum Pseudo Labeling (Zhang et al., 2021)

FlexMatch builds upon FixMatch by adding curriculum pseudo-labeling, which seeks to select unlabeled images for pseudo-labeling in an optimal order. To do so, an adaptive threshold is created to allow classes which are infrequent in the unlabeled set to be selected more often. To gauge this frequency, the authors calculate the "learning effect", which counts the number of unlabeled examples whose maximum class predictions are

greater than a threshold for each given class, as shown in Equation 4.

$$\sigma_t(c) = \sum_{n=1}^{N} \mathbb{1}(\max(p_{m,t}(y|u_n)) > \tau) \cdot \mathbb{1}(\arg\max(p_{m,t}(y|u_n) = c) \quad \text{(Equation 4)}$$

This number is higher for classes which occur frequently and at higher confidences. The learning effect is then normalized by the maximum learning effect across all classes in Equation 5.

$$\beta_t(c) = \frac{\sigma_t(c)}{\max_c \sigma_t} \quad \text{(Equation 5)}$$

The normalized learning effect per class, $\beta_t(c)$, then creates a new class-specific scaled threshold, which lowers the confidence threshold for less common classes, shown in Equation 6.

$$\mathcal{T}_t(c) = \beta_t(c) \cdot \tau \quad \text{(Equation 6)}$$

Thus, more images of rarer classes in the unlabeled dataset are selected for pseudo-labeling. With respect to transferring this technique to the segmentation context, the strong augmentation limitations from FixMatch persist. Furthermore, the class balancing method would need to be altered for segmentation datasets, which have multiple classes per image.

In Defense of Pseudo-Labeling: An Uncertainty Aware Pseudo Label Selection Framework for Semi-Supervised Learning (Rizve et al., 2020)

In Rizve et al. (2020), the authors train a model on a labeled set and then use the model to predict class probabilities on an unlabeled set. They calculate the standard deviation of the model predictions over 10 forward passes using Monte Carlo dropout. This standard deviation functions as the uncertainty score. Images with maximum class predictions greater than a threshold and uncertainty scores less than a threshold value are then used to train the model. These thresholds are found empirically by a validation dataset. To use this in the segmentation context, the uncertainty score and class prediction would need to be adapted since each pixel has class predictions. In contrast, in the image classification problem, each image has a class prediction.

Uncertainty-Aware Deep Co-training for Semi-supervised Medical Image Segmentation (Zheng et al., 2021)

To understand Zheng et al. (2021), I start with the co-training concept, which was introduced by Blum and Mitchell (1998). Co-training trains classifiers from multiple views of the same scene. This creates viewpoint diversity and results in a more effective ensemble.

In 2018, Qiao et al. extended the co-training concept to image classification by using

adversarial examples to enforce viewpoint diversity. Adversarial examples, briefly, are input examples that are altered via knowledge of the model's gradient in such a way as to cause the model to incorrectly classify the image. Figure 4.4 shows how each model is trained on the adversarial example of the other model, which creates a more robust ensemble.

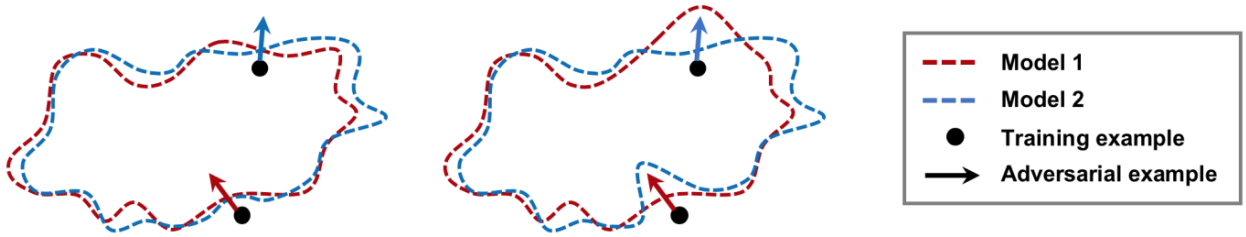

Figure 4.4: This figure shows the effect of adversarial training from Qiao et al. (2018).

Peng et al. (2020) then extend this co-training framework to the segmentation context. They focus on three losses: one that uses the standard cross-entropy loss on the labeled data, an ensemble agreement loss on the unlabeled data, and the diversity loss on both the labeled and unlabeled data.

Zheng et al. (2021) build upon Peng et al. (2020), but add an uncertainty weight to the supervised and unsupervised losses. The uncertainty is calculated per image for each model by 10 stochastic forward passes using Monte Carlo dropout and taking the entropy of the averaged output per pixel. Entropy is a measure of information spread, so if the average output probability of the maximal class is low, then the model is not confident about the prediction. Therefore the uncertainty of the pixel is higher. The uncertainty scores are used differently for the supervised and unsupervised losses. In the supervised loss, the cross entropy loss of each pixel is weighted by that pixel's uncertainty score for each model, resulting in the model learning more from the pixels with higher uncertainty, since I have certainty of their labels. Figure 4.5 from Zheng et al. illustrates the supervised learning process guided by uncertainty.

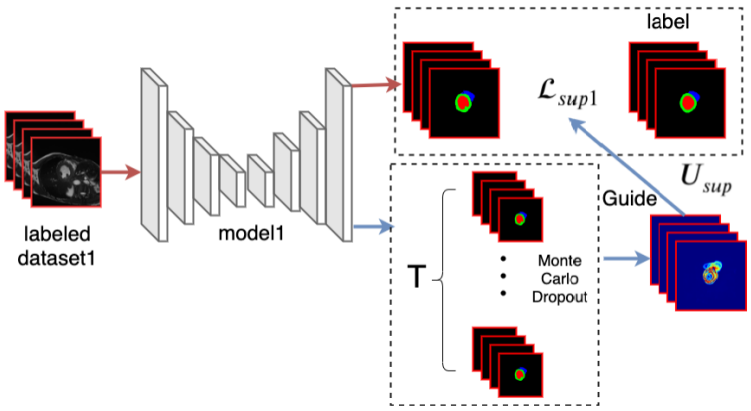

Figure 4.5: Supervised Learning with Uncertainty from Zheng et al. (2021).

For the unsupervised data, the uncertainty weights from each model are averaged, scaled, and multiplied by -1, which results in learning more from the pseudo-labels from

pixels about which the models are jointly certain. Figure 4.6 from Zheng et al. shows both models being trained using their joint uncertainty on the unlabeled data.

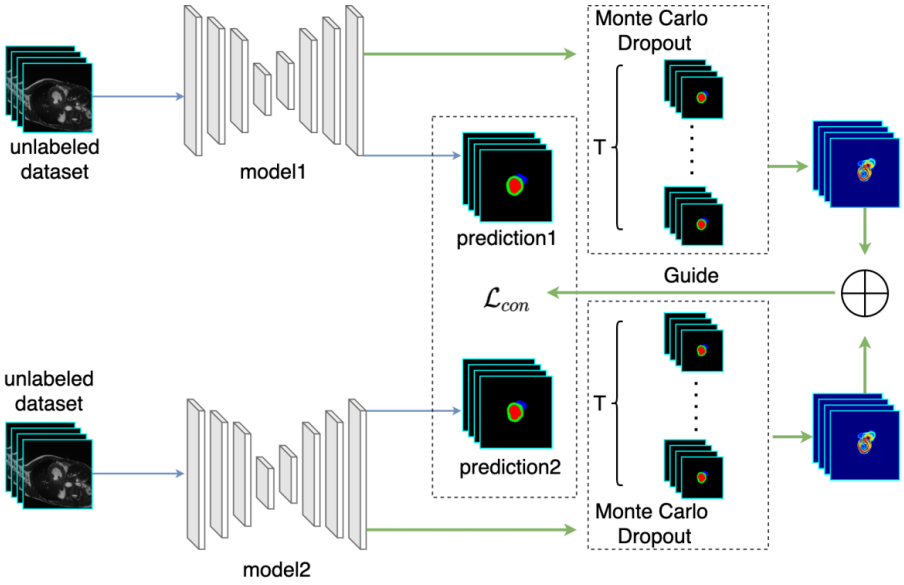

Figure 4.6: Unsupervised Learning with Uncertainty from Zheng et al. (2021).

One issue with this approach is that all unlabeled images are used, but only certain pixels are discounted according to the model uncertainty. Thus, this method does not address how to downselect to a useful subset when the unlabeled data set is too large to be feasibly used.

4.3 Discussion and Limitations of Literature

The FixMatch algorithm (Sohn et al., 2020) and the FlexMatch algorithm (Zhang et al., 2021) use unlabeled data in a consistency regularization setup. FlexMatch additionally has a threshold that adapts to under-observed classes in the multiclass classification problems. Rizve et al. (2020) used uncertainty quantification, with a deterministic threshold to select the images with the most confidence and lowest uncertainty from the unlabeled set. While these algorithms provide selection criteria on an image-level basis, they are not adaptive in binary settings nor applied to segmentation problems. All three of these algorithms concern classification problems. While Zheng et al. (2021) address segmentation problems, it sets no threshold for selecting unlabeled images. Additionally, the uncertainty quantification for weighting the loss function is expensive to calculate for every iteration in an epoch. Furthermore, none of these metrics describe a process to verify the trustworthiness of an uncertainty metric.

In light of these limitations, this section seeks to answer the following: How can unlabeled data be incorporated and trusted for image-level selection?

The contribution of this chapter answers this question by adapting confidence and uncertainty quantification methods from classification to segmentation setting for image-level selection. I also use the correlation of the uncertainty metric with the dice score

on the training set to verify trustworthiness for the uncertainty quantification. The impact of this contribution is to enable the prioritization of images that maximize performance and provide trust in the model via intuitive visualization of uncertainty.

4.4 Pre-processing of Datasets

First, I discuss the preprocessing of 3 datasets derived from whole-slide-images in this evaluation. The eosinophilic esophagitis (EoE) labeled dataset consists of 514 images at 512 x 512 pixels. Each image is used as it is with its binary mask for eosinophil pixel annotations. The CoNSeP dataset consists of 41 images, each with a size of 1000 x 1000 pixels. The images are sliced into 500 x 500 pixel sub-image patches with no overlap, resulting in 164 images. To simplify the problem, all nuclei are collapsed into a single type. The MoNuSeg dataset consists of 44 images, each with a size of 1000 x 1000 pixels. Each image is sliced into 500 x 500 pixels sub-images with no overlap, resulting in 176 images. Each image has pixel-wise nuclei annotations.

4.5 Methodology

4.5.1 Examine Correlations

The correlations between image-level model performance on the labeled dataset, referenced in Table 4.1 as the training dice, and the associated uncertainty and confidence measures are examined. 2 different types of uncertainty measures are used: 1) the standard deviation uncertainty (Gal and Ghahramani 2016; Rizve et al., 2020), and 2) entropy (Shannon 1948). The standard deviation uncertainty is the standard deviation of the output probabilities for 20 forward inferences through a U-Net model using Monte Carlo (MC) dropout. For entropy, the output probabilities of the MC dropout models are averaged, and then the entropy per pixel is calculated. Since each pixel is assigned uncertainty and confidence values, the values are averaged across all pixels per image to achieve a single image-level value, and they are plotted against the model performance per image. The central assumption in using confidence and uncertainty in pseudo-label selection is that as confidence increases and uncertainty decreases, the quality of the pseudo-labels should also increase. Investigating these correlations allows us to assess the robustness of this assumption for image-level confidence and uncertainty values for the segmentation task. Moreover, a model which violates this assumption would not be able to differentiate between pseudo-labels of varying quality. Also, these correlations cannot be examined for the unlabeled dataset, because, by definition, there are no true labels for them to use to calculate the dice coefficient. Thus, the magnitude and direction of these correlations for the labeled dataset make the best proxy for how the model would behave on the unlabeled data, indicating the model's usefulness in pseudo-label image selection.

4.5.2 Pseudo-Label Selection via Thresholds and Weighting

In line with Rizve et al. (2020), I set the minimum confidence and maximum uncertainty thresholds to select which unlabeled images to pseudo-label and use for training for each dataset. Additionally, I propose adaptive thresholds that are determined by a given model's baseline uncertainty, this is because the distribution of uncertainty values varies per model, as shown in Figure 4.8. For the fixed thresholds, empirically select values by looking at the histogram of the image-level uncertainty and confidence values produced by 8 different random seeds per dataset. Figure 4.7 shows the confidence histogram for the images in MoNuSeg.

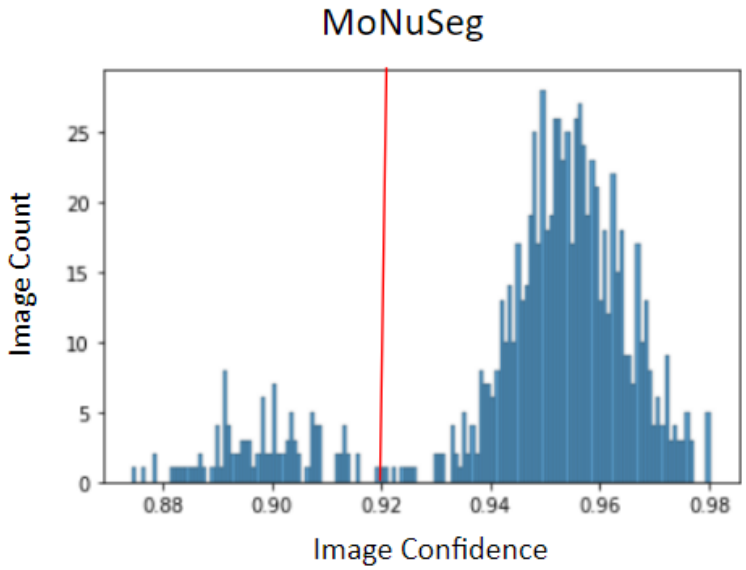

Figure 4.7: This figure shows the histogram of the model's confidence for each image for the MoNuSeg dataset.

The confidence, standard deviation, and entropy thresholds are 0.99, 0.02, and 0.04 for EoE, 0.75, 0.135, and 0.55 for CoNSeP, and 0.92, 0.09, and 0.14 for MoNuSeg, respectively. In Table 4.1, the thresholding techniques are referred to as "Conf τ StD τ" for confidence and standard deviation and "Conf τ E τ" for confidence and entropy. The adaptive thresholds are: minimum 25% percent entropy and sigmoid-weighted entropy.For minimum 25% entropy, I select the images with entropy values in the lowest 25% of the unlabeled set for pseudo-labeling. These images are more likely to be helpful because the model has greater certainty about them. The other 75% of the unlabeled set is not used for pseudo-labeling. Figure 4.8 shows this threshold for 2 different models. For the sigmoid weighted entropy, the weight of each image in the loss function during training is set by $S(x) = \frac{1}{1+e^{-x}}$, where $x = \alpha(H(i) - \mu_{H(lab)})r_{lab}$, α is a tunable scaling hyperparameter set empirically to 1000, $H(i)$ is the image-level entropy for image i, $\mu_{H(lab)}$ is the mean entropy value of the labeled dataset and r_{lab} is the correlation of the image-level entropy and with the dice coefficient of the labeled dataset. The sigmoid function assigns larger weights to images with lower entropy values, focusing the model on pseudo-labels with higher

expected accuracy. This weighting has the effect of creating an adaptive, gradual threshold that weights each image's pseudo-labels in relation to the model's certainty about the image and in relation to the uncertainty of the trained dataset. I compare the results of using these methods to not using pseudo-labels and using all pseudo-labels with equal weights in the loss function during training.

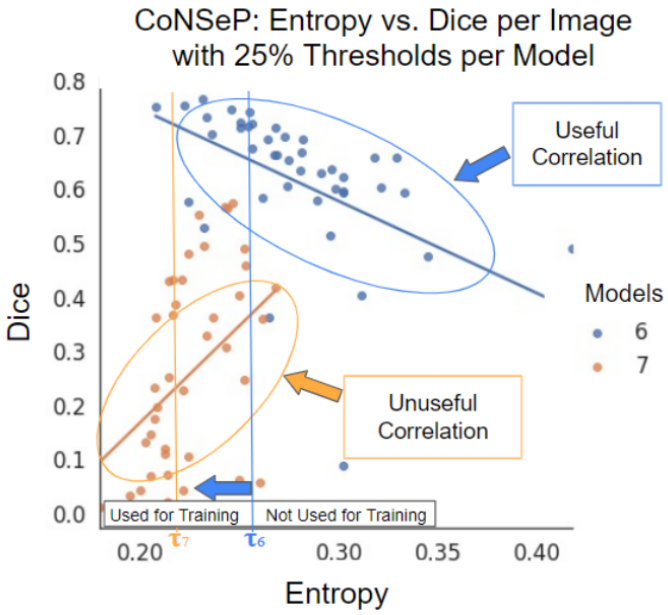

Figure 4.8: This figure shows the correlation of entropy and dice from 2 different models trained on the CoNSeP dataset. The threshold line splits the 25% lower entropy values.

4.5.3 Model Setup and Training

A U-Net architecture was used with randomly initialized weights for each seed per dataset. The Adam optimizer and a learning rate of 2e-4 were used. Each model was trained on half of the training data and the other half was used as the unlabeled set for pseudo-labeling. For the initial labeled training, I used early stopping of 12 epochs without improvement in the validation loss, up to 100 epochs total. After this, uncertainty and confidence values were quantified for each unlabeled image. Then, the relevant images were selected for pseudo-labeling and assigned weights for the loss function for each technique mentioned. The model was trained for 3 epochs using labeled and pseudo-labeled data, where the model with the best validation accuracy was saved. No pseudo-labels were included in the validation set. The process repeated until the model performance on the validation set did not increase for 3 rounds of training on pseudo-labels. The converged model was then evaluated against a hold-out test set.

4.6 Experiments and Results

I first examine the correlations of the image-level confidence and uncertainty values with the model performance per image. This is realized as the dice coefficient of the pseudo-labels and the true labels using 8 randomly initialized models on the labeled

dataset. Each model's dice coefficient against a test set is also evaluated. Additionally, the expected calibration error (ECE) (Naeini et al., 2015) is used to determine whether model calibration corresponds to changes in the correlations or performance. Each model for a given dataset has the same validation set, but the labeled and unlabeled datasets are randomly partitioned for each seed and are equal in size.

4.6.1 Correlation Results

For the EoE dataset in Table 4.1, the image-level uncertainty values for images in seeds 4 and 6 have a positive or zero correlation with the image-level training dice coefficients. Also, the image-level confidence values for these seeds have a negative correlation with dice coefficients. This signifies that as the uncertainty for a given image decreases and its confidence increases, the model's performance on that image decreases. This is counter-intuitive since the assumption of pseudo-labeling is that as uncertainty decreases and confidence increases, the pseudo-labels' usefulness in an image, measured here by the dice coefficient between the pseudo-labels and the true labels, should increase. This correlation, though in the minority of the randomly initialized runs, does demonstrate an instance where the image-level confidence and uncertainty values from a model are not useful in selecting pseudo-labels, since minimizing image-level uncertainty results in selecting images with less accurate pseudo-labels. Furthermore, the models with these unexpected correlations do not present a consistently poor test dice coefficient or low ECE compared to those with the expected correlations. Thus, neither poor initialization nor calibration accounts for this correlation phenomenon.

For the CoNSeP dataset in Table 4.1, the image-level confidence and entropy values for images in seeds 2, 3, 4, 5, and 7 have negative and positive correlations, respectively, with the image-level training dice values. Again, these correlations undermine the central assumption in selecting pseudo-labels. However, these seeds have a noticeably lower test dice coefficient than those with the expected correlations, seeds 0, 1, and 6. Welch's t-test confirms with a p-value of 0.007 that the average test dice coefficient of the seeds with the expected correlation is higher than those without it. The ECE values do not account for the correlation or performance differences. For the MoNuSeg dataset in Table 4.1, the image-level entropy values for seeds 0 and 2 have a positive correlation with the image-level training dice values. Again, this indicates that as uncertainty increases the accuracy of the pseudo-labels increases as well. The image-level confidence values for seed 0 correlate negatively with image-level training performance. The test performance for seed 0 is low, which is statistically significant according to a t-test comparing it to the other models' test dice scores. The test dice score is not abnormal for seed 2, possibly due to its positive confidence correlation with the dice score. The model for seed 2 has the lowest ECE value, and the ECE values for other seeds are closely grouped. Thus, poor model calibration does not account for these correlation or performance differences.

4.6.2 Pseudo-Label Selection Results

Table 4.2 shows that most pseudo-label selection methods give a statistically significant boost to the model's test performance for EoE, but this occurs for all models, regardless of whether the correlation is as expected. Table 4.2 shows that no pseudo-label

selection method gives a statistically significant boost to the model's test performance for the CoNSeP dataset. This holds for the models with the expected correlations, seeds 0, 1, and 6, and the poorly correlated models, seeds 2, 3, 4, 5, and 7. For the minimum 25% entropy pseudo-label method, the seeds with the expected correlations did not differ significantly from the average dice coefficients of the models before pseudo-labeling. However, the seeds that were poorly correlated did have a lower dice performance on the test set that was statistically significant.

Table 4.2 shows that no pseudo-label selection method gives a statistically significant boost to the model's test performance for the MoNuSeg dataset. This holds for the models with the expected correlations, seeds 1, 3-7, and the poorly correlated models, seeds 0 and 2. The pseudo-labels worsen the training for the models, though only so in a statistically significant way for "same" and "Conf τ StD τ."

4.7 Discussion

Across the three datasets, randomly initialized models did not always have the expected correlations of image-level confidence and uncertainty with image-level performance. This contradicts an intuitive assumption of pseudo-labels: as model certainty increases, the quality of the pseudo-labels should also increase. It is possible that the differences in these metrics are averaged out when the entire image is taken into account. Furthermore, models for the EoE dataset increased performance when trained on pseudo-labels, regardless of the individual model's correlations. This may be due in part to the sparsity of the EoE dataset. Since very few Eos are contained in each image, the model may be too conservative due to class imbalance. Pseudo-labels, even if not always correct, may help with addressing the bias in the data of a lack of Eos. Though not robust, the correlations did prove informative in certain contexts. For MoNuSeg and CoNSeP, the models where the confidence correlated negatively and the entropy correlated positively with the training dice performed worse than those models with the opposite correlations in a manner that was statistically significant. This indicates that if a model has certainty metrics that do not correlate with the performance as expected, it may have worse performance on the test set than a model that has the expected correlations. For CoNSeP, using the minimum 25% entropy pseudo-labels worsened the model performance for the poorly trained seeds, whereas the models with the expected correlations were not negatively impacted in a significant way by any of the pseudo-label weightings.

4.8 Conclusion

Before using pseudo-labeling, examining the correlations between certainty metrics and a model's performance on labeled imagery may help indicate whether a model's uncertainty about an image is informative. Thus, a positive correlation for confidence and a negative correlation for uncertainty are necessary but insufficient indicators of the utility of pseudo-labeling for a given model. Furthermore, averaging certainty metrics across an image may obfuscate the complexity inherent in segmentation, particularly when there is a class imbalance. Further work examining these correlations on a pixel-wise basis would be insightful.

Table 4.1 This table shows the correlation of the confidence, the standard deviation of MC dropout predictions, and the entropy with dice coefficient on the training set, as well as the test dice coefficient and expected calibration error, for multiple model runs across the EoE, CoNSeP, and MoNuSeg datasets.

EoE					
	Corr. with Training Dice				
Seed	Conf	StDev	Entropy	Test Dice	ECE
0	0.032	-0.053	-0.042	0.477	0.010
1	0.06	-0.075	-0.064	0.46	0.011
2	0.057	-0.066	-0.06	0.445	0.012
3	0.029	-0.024	-0.083	0.457	0.009
4	**-0.06**	**0.034**	**0.069**	0.433	0.011
5	0.045	-0.053	-0.043	0.46	0.010
6	**-0.015**	**-0.002**	**0.014**	0.485	0.011
7	0.18	-0.21	-0.183	0.49	0.011
CoNSeP					
	Corr. with Training Dice				
Seed	Conf	StDev	Entropy	Test Dice	ECE
0	0.091	-0.131	-0.145	0.514	0.197
1	0.267	-0.074	-0.535	0.599	0.104
2	**-0.447**	-0.29	**0.442**	0.283	0.171
3	**-0.064**	-0.042	**0.011**	0.275	0.183
4	**-0.159**	-0.142	**0.134**	0.261	0.147
5	**-0.314**	-0.272	**0.419**	0.291	0.343
6	0.304	-0.125	-0.47	0.615	0.077
7	**-0.382**	**0.229**	**0.421**	0.318	0.727
MoNuSeg					
	Corr. with Training Dice				
Seed	Conf	StDev	Entropy	Test Dice	ECE
0	**-0.043**	-0.072	**0.007**	0.625	0.117
1	0.219	-0.35	-0.272	0.725	0.101
2	0.146	-0.376	**0.025**	0.71	0.075
3	0.363	-0.387	-0.439	0.74	0.086
4	0.087	-0.185	-0.13	0.74	0.114
5	0.266	-0.285	-0.345	0.699	0.144
6	0.149	-0.249	-0.172	0.75	0.124
7	0.246	-0.293	-0.281	0.761	0.089

Table 4.2 Pseudo-Label vs No Pseudo-Label with Welch's t-test: This table gives the test dice coefficients for the EoE, CoNSeP, and MoNuSeg datasets averaged over multiple runs across multiple types of pseudo-label selection.

Pseudo-Label Selection	**Test Set Dice per Dataset**		
	EoE	**CoNSeP**	**MoNuSeg**
No Pseudo-Label	0.46 ± 0.02	0.39 ± 0.15	0.72 ± 0.04
All Data[a]	0.60	0.72	0.76
Same	$\underline{0.52 \pm 0.03}$	0.43 ± 0.15	$\underline{0.66 \pm 0.04}$
Sigmoid Entropy	$\underline{0.5 \pm 0.05}$	0.38 ± 0.14	0.6 ± 0.09
Min. 25% Entropy	$\underline{0.54 \pm 0.03}$	0.36 ± 0.16	0.66 ± 0.07
Conf τ StD τ	$\underline{0.51 \pm 0.04}$	0.36 ± 0.13	$\underline{0.66 \pm 0.05}$
Conf τ E τ	$\underline{0.51 \pm 0.03}$	0.35 ± 0.12	0.69 ± 0.03

[a]This signifies the upper limit.

*The underlined numbers are different from the No Pseudo-Label dice, and the difference is statistically significant, determined by Welch's T Test with a p-value of < 0.05.

Chapter 5

Identification and Effect of Bias in Histopathological Segmentation

5.1 Summary

Segmentation algorithms in histopathology provide a basis for scalable disease quantification and classification. Labeled data guides the training of these algorithms to learn a feature representation of the cells of interest. However, labeled training data is not always available for all patients. This leads to poor performance on patients with only unlabeled data. Semi-supervised learning (SSL) is used to tackle this scenario where labeled patient data is leveraged with unlabeled target patient data to improve the model performance. Even so, SSL techniques may fail to generalize to unlabeled patient data due to shifts in the patient's histological features. In this chapter, I apply an unsupervised clustering technique to provide a distributional understanding of multiple patients' tissue features. This clustering enables the identification of different types of biases that can occur in histopathological datasets. These biases have interpretable clinical action steps that I address.

5.2 Literature Review

The incorporation of unlabeled data concerns not only the image-level selection and trustworthiness of the model's uncertainty quantification but also the identification of bias between the labeled and unlabeled sets. This bias can prevent models from generalizing well to unseen data. In the following sections, I describe the existing approaches to identifying differences in the feature distribution of labeled and unlabeled sets and the semi-supervised algorithms I use.

5.2.1 Distribution Mismatch and Bias

Semi-supervised learning (SSL) techniques encounter issues when the assumption that the labeled and unlabeled sets are drawn independently from identical distributions is violated (Calderon-Ramirez et al., 2023). These issues occur because differences between labeled and unlabeled data sets can lower SSL model performance, as demonstrated in image classification tasks in Oliver et al. (2018) and Calderon-Ramirez et al. (2022). Existing techniques to address the distributional mismatch between labeled and unlabeled sets in SSL identify when the unlabeled set contains data that are out-of-distribution (OOD) (Calderon-Ramirez et al., 2022), quantify the dataset dissimilarity (Calderon-Ramirez et al., 2023), and aim to lessen any negative effect on model performance (Kurian et al., 2023).

Specifically within histopathology, mismatch in dataset distribution manifests as various types of biases. Burkhardt et al. (2011) note the inherent presence of sampling bias in tissue selection, despite the standardization of sampling techniques to reduce variability (Bucci, 2002). Institutional bias also plays a role, where a model learns features specific to the image's source, such as Hospital A or Hospital B. Learning these differentiating features, which are not medically relevant, can interfere with deep learning applications (Dehkharghanian et al., 2023). Many works seek to reduce that bias via stain normalization (Bejnordi et al., 2016; Ciompi et al., 2017) and color augmentation (Lafarge et al., 2017; Lin et al., 2018), as noted in Komura and Ishikawa (2018). Bigdoli et al. (2022) also minimize institutional bias but do so through an evolutionary feature selection algorithm to select deep learning features of histopathological significance that do not contribute to institutional differentiation. Hägele et al. (2020) used heatmaps to visualize different types of bias in histopathological images. In the following section, I examine the methods in the literature.

Dealing with distribution mismatch in semi-supervised deep learning for COVID-19 detection using chest X-ray images: A novel approach using feature densities (Calderon-Ramirez et al., 2022)

Calderon-Ramirez et al. (2022) uses labeled and unlabeled X-ray images from various clinics to detect COVID-19 with the MixMatch algorithm. They analyze the impact of mismatches in the labeled and unlabeled distributions, and they contribute 2 methods to score data as coming from the labeled distribution. They use these scores to filter unlabeled data, which may harm the performance of an SSL algorithm. To create the scores, they pass each image through a classifier pre-trained on ImageNet and extract the feature embedding. Then, for each element in the n' dimensional feature embedding, a density function is approximated via a set of normalized histograms. Each element in the set is a density function. Given an unlabeled image, the feature embedding is created similarly via the pre-trained CNN. Then, the probability of each element in the feature embedding of the unlabeled image is found using the corresponding histograms created by the labeled set. The negative log-likelihood of the product of each of the unlabeled features' probabilities is calculated to give the score. For the second score, a Gaussian distribution is assumed over the features, and the Mahalanobis distance between the unlabeled image embedding and the mean values and covariance from the labeled set is calculated. Unlabeled data with high scores, signifying greater differences with the labeled feature set, are discarded. Thus, these scores do provide a way to filter out unlabeled images that present a large covariate shift. However, the nature or cause of the difference is not further examined. Also, this method would likely have trouble in sparsely labeled datasets that failed to produce density functions corresponding to the full range and complexity of naturally occurring features.

Dataset Similarity to Assess Semi-Supervised Learning Under Distribution Mismatch Between the Labeled and Unlabeled Datasets (Calderon-Ramirez et al., 2023)

Calderon-Ramirez et al. (2023) builds dataset dissimilarity measures to select an optimal subset from the unlabeled set for maximizing performance in the SSL Mixmatch algorithm for evaluation on MNIST, CIFAR-10, and Fashion MNIST. They use a feature embedding via a pre-trained CNN as in Calderon-Ramirez et al. (2022) and create four measures between subsets of the labeled and unlabeled sets. Two of the distances are based on the Euclidean and Manhattan distances between the feature embeddings of samples from the labeled and unlabeled sets. The other 2 measures are based on the Jensen–Shannon and cosine distance and measure the divergence in probability densities between the histograms derived from the feature embeddings in the labeled and unlabeled datasets. They show that the unlabeled subsets with the smallest distance to the labeled set correlate with higher accuracy when used in the MixMatch algorithm.

These scores provide a way to select an optimal subset of unlabeled images for use in SSL algorithms. However, as in Calderon-Ramirez et al. (2022), the nature or cause of the difference between the subsets is not further examined, nor are the sparsity concerns addressed.

Robust Semi-Supervised Learning for Histopathology Images through Self-Supervision Guided Out-of-Distribution Scoring (Kurian et al., 2023)

Kurian et al. (2023) creates a score to detect samples that are out-of-distribution for SSL algorithms using histopathology images. First, they use the self-supervised framework SimCLR (Chen et al., 2020), which learns the underlying structure of all the labeled and unlabeled data via consistency regularization and a contrastive loss. They then use the latent features from the model trained via SimCLR from both datasets to create a Gaussian Mixture Model (Reynolds, 2009). Each Gaussian is assigned an impurity score, which describes the probability of labeled samples in a cluster relative to the probability of all samples in a cluster, with the idea that a cluster with fewer labeled samples has outlier data. An out-of-distribution score is calculated for each unlabeled sample by a summation of the product of the cluster impurity and the posterior probability of the sample belonging to the cluster over all clusters. In training an SSL algorithm, a cluster is chosen with probability inversely proportional to its impurity score. Then, the unlabeled data are sampled from that cluster with probability inversely proportional to their OOD scores. In this way, the method prioritizes unlabeled samples that are less likely to be out-of-distribution. Figure 5.1 below shows the full process from Kurian et al. (2023).

This approach allows for sampling unlabeled data in SSL that probabilistically prioritizes data from feature distributions similar to the labeled set. However, the clusters are not examined for any bias they might reveal. Furthermore, this is applied to a classification problem and does not address the segmentation context or sparse segmentation labels. The contrastive learning setup is an additional expensive step that may be unnecessary, given the prevalence of pre-trained networks.

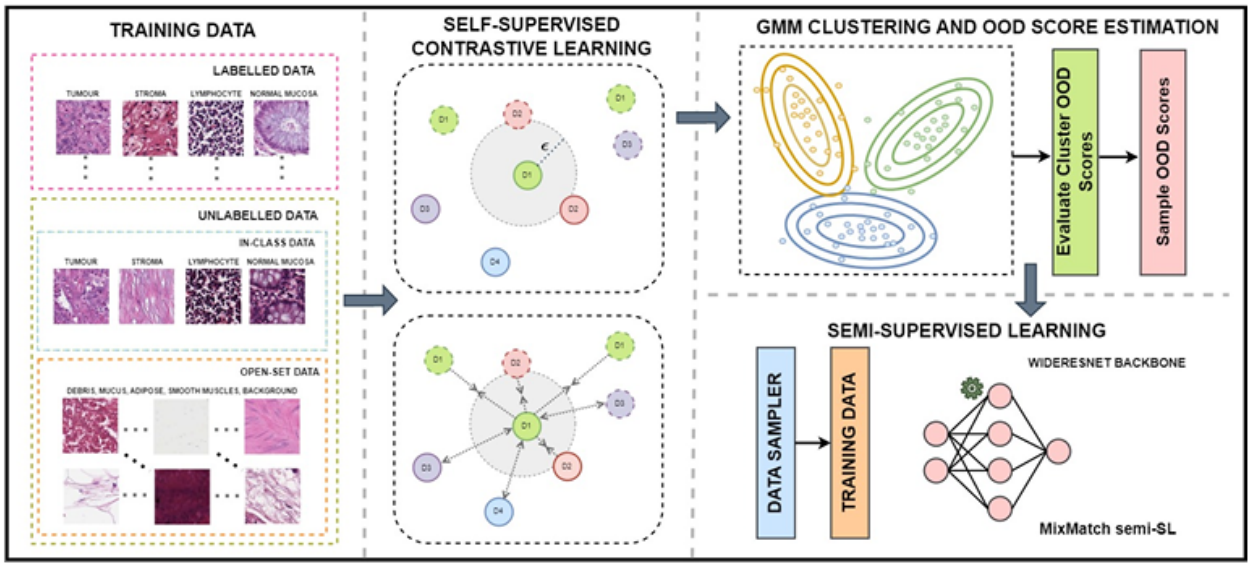

Figure 5.1: The figure shows the process from Kurian et al. (2023) of using self-supervised contrastive learning to train a classifier on the labeled and unlabeled data, projecting the unlabeled data into an embedding space from that classifier, performing GMM clustering and OOD score estimation, and then using those values to probabilistically unlabeled data with features more similar to the labeled set.

Bias Reduction in Representation of Histopathology Images using Deep Feature Selection (Bidgoli et al., 2022)

Bidgoli et al. (2022) starts with feature vectors from the embedding of histopathology patches passed through two networks; KimiaNet, pre-trained on histopathology images and DenseNet-121, pre-trained on ImageNet. The feature vectors are then sub-selected via an evolutionary algorithm to maximize the image search quality, minimize the number of features, and minimize institutional bias. While this method may reduce the dimensionality of the embeddings and the effect of the bias, it does not reveal the cause. Rather, the institutional bias in the only one assumed to occur.

Resolving challenges in deep learning-based analyses of histopathological images using explanation methods (Hägele et al., 2020)

Hägele et al. (2020) visualize the heatmaps of CNNs to detect bias. The heatmaps are produced via backpropagation of the activated neurons to the input pixels with respect to their contribution to the classification. In examining the heatmaps, they reveal and examine three kinds of bias: dataset bias, class-correlated bias, and sampling bias. Dataset bias concerns non-medically relevant features which persist throughout the dataset. An example they give is when the feature determining the classification always appears in the same location in the image. The heatmap reveals that a model trained on a dataset with this bias focuses on the center of the image for the classification, even when the distinguishing feature is not in the center in the test set example. The class-correlated bias is similar in that it is an overfitting error, where the model learns features that co-occur with the class of interest even though they are unrelated to the classification. Sampling bias deals with

issues such as not sampling a representative sample of features for training. For all three biases, the heatmaps are used per patch to identify them. Since the heatmaps relate to the activation backpropagation of the activated neurons for classification, this method makes the most sense in the classification context. This process would not translate as clearly to the segmentation context because each pixel receives a class label rather than the image as a whole. Additionally, the visual inspection of the heatmaps requires aggregation and analysis of the dataset as a whole to establish common patterns, which may be subjective in interpretation if the feature is not prevalent throughout the dataset.

5.2.2 Stain Normalization and Color Augmentation

Stain normalization and color augmentation techniques minimize the effect of institutional biases but do not analyze their root causes or characterize them. Lafarge et al. (2017) use adversarial training to encourage the model to learn a domain invariant space that does not allow for differentiation of the institutional source for mitosis detection in breast cancer histopathology images. Ciompi et al. (2017) show the importance of stain normalization for classification accuracy of colorectal cancer in histopathology images. More generally, if tissue is sampled in a region with a much lower probability of the class of interest occurring, then the sparsity of the data will remain constant, regardless of how much it is augmented or normalized. The signal of the class of interest would still not be present because it was not sampled, so its feature distribution would still evade thorough characterization. Thus, they do not address the prior probability shift that occurs in sampling bias. Next, I describe an SSL algorithm that is adapted in this chapter for use in the bias analysis.

Semi-Supervised Semantic Segmentation with Cross Pseudo Supervision (Chen et al., 2021)

Chen et al. (2021) train two models on the output of the other on unlabeled data. That is, each model is trained on the pseudo-labels of the other model on the unlabeled and unlabeled data, along with each model's supervised loss, as shown in Figure 5.2. This method combines a form of consistency regularization and supervised learning, and I build on it for this work.

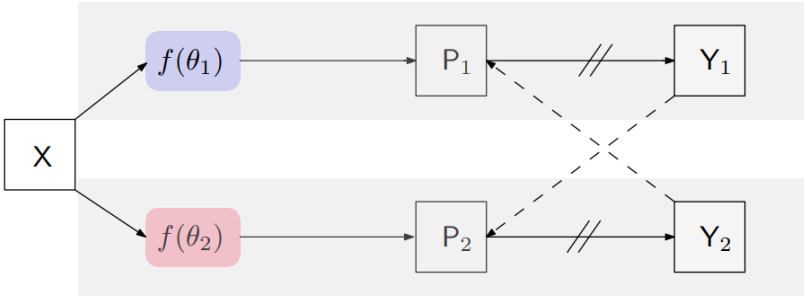

Figure 5.2: Cross-Pseudo Supervision from Chen et al. (2021).

5.3 Discussion and Limitations of Literature

While individual data points are filtered via an out-of-distribution score to improve SSL performance in Calderon-Ramirez et al. (2022), the bias itself is not identified or characterized. Similarly, the dataset dissimilarity measures in Calderon-Ramirez et al. (2023) enable the selection of an optimal subset from the unlabeled set for maximizing SSL performance. However, the cause of the difference in the subsets is not further examined. Both works do not account for sparse feature representation in limited labeled settings.

Kurian et al. (2023) present a sampling method that draws unlabeled samples from distributions that have features similar to the labeled set, but the clusters themselves are not examined for what kind of bias exists. The sampling process depends on the feature distribution similarity of the labeled and unlabeled sets. This is problematic for the per-patient approach described in my methodology because it would require the cluster parameters to be calculated for each patient analyzed, greatly increasing the complexity of the analysis. Additionally, contrastive learning pre-training is an additional costly step, likely unnecessary given the availability of pre-trained networks.

In their dimensionality reduction method, Bidgoli et al. (2022) only address institutional bias. Hägele et al. (2020) visualize the heatmaps of CNNs to detect bias, but this requires aggregation and analysis that may be difficult to interpret. The heatmaps only apply in classification settings and are not clearly extensible to segmentation problems. Stain normalization and color augmentation techniques minimize the effect of institutional biases but do not analyze their root causes or characterize them. Nor do they address the prior probability shift that occurs in sampling bias.

Overall, these methods focus on mitigating the effects of institutional bias and outlier features, but they do not produce a distributional understanding of the features that characterize the biases. Furthermore, none consider the setting of withholding a single patient's labeled data while training on that patient's unlabeled data and the other patients' labeled data. Such a distributional understanding would help reveal the bias inherent not only with respect to the institution but due to the individual patient's data.

In light of these limitations, this section seeks to answer the following: How can biases inherent in unlabeled and labeled data be identified that would hinder the generalization of semi-supervised algorithms?

The contribution of this chapter answers this question by clustering the unlabeled and labeled patients' data in a classifier's embedding space and using the distribution of each patient's images across the clusters as a means for understanding when the SSL models will not generalize well on a given patient's data. I then show how clustering the unlabeled and labeled data reveals the different types of bias present, specifically sampling bias and labeling bias. The impact of this contribution gives clear interpretability about biases that enables the correct clinical solution, reducing cost and minimizing procedures necessary for patients.

5.4 Pre-processing of Datasets

I examine two datasets in this chapter, the EoE and CD datasets. Each dataset has patches of 512 x 512 pixels at 40x magnification. I consider all labeled and unlabeled patches in both datasets. The EoE dataset consists of 514 labeled and 240,526 unlabeled images, and the CD dataset contains 200 labeled images and 291,779 unlabeled images. Each labeled image has a binary mask for eosinophil pixel annotations.

5.5 Methodology

I examine the performance of semi-supervised learning (SSL) algorithms to segment eosinophils in 2 datasets on a per-patient basis. I train a model on $N-1$ patients' labeled data and the unlabeled data from the Nth patient, then test on the Nth patient's labeled data. For certain patients, the SSL methods struggle to increase performance relative to a supervised baseline. I hypothesize this is due to a shift in the patients' feature distributions. Instead of casting such a shift as a domain adaptation problem, I seek to provide interpretability behind the feature shifts. Specifically, a Gaussian Mixture Model is learned from the unlabeled images of all patients to cluster each image by tissue type. My analysis of the clustering results reveals biases in the data for each patient for which the SSL methods do not increase performance. Furthermore, each bias has a clinically interpretable solution.

5.5.1 Experimental Setup on a Patient-by-Patient Basis

I consider bias on a patient-by-patient basis. This setup mimics real-world settings, where the data is first digitized before it is labeled, if at all. Also, I want to identify the clinical action step required to maximize segmentation performance for the individual patient. For example, another biopsy may be required if sampling bias is detected in a patient's data. Conversely, if no bias is detected, the model is more likely to generalize well without further invasive procedures.

This paradigm of patient-based bias has not been explored in the literature. Instead, the focus has been on minimizing the effect of institutional bias or images with out-of-distribution features. I show bias can also exist per patient, where the clinically interpretable solution applies to a particular patient.

5.5.2 Semi-Supervised Segmentation Approaches

I found that the basic CPS setup did not perform well on my datasets, so I augmented it. I build upon CPS by specifying a balanced batch sampling process and by selecting unlabeled data that are more likely to benefit from pseudo-labeling, which I call CPS-Adapted (CPSA). Furthermore, I incorporate image-level confidence estimation in weighting the loss for the unlabeled data in Equation 8, inspired by Xie et al. (2022). I also incorporate data augmentations optimized for histology (Tellez et al., 2019), such as color jitter and leaving out greyscale.

$$C_x = \frac{1}{|WxH|} \sum_{i \in x} 1(p_i > \tau) \qquad \text{(Equation 8)}$$

p_i is the maximum class probability for pixel i. τ is a deterministic threshold, where τ = 0.968 in Xie et al. (2022), which I also find sufficient for my experiments. Next, I define the CPSA loss with confidence weighting in Equation 9.

$$L_{csps,conf} = \frac{1}{|B|} \sum_{x \in D_u}^{B} \frac{C_x}{WxH} \sum_{i=1}^{WxH} l_{ce}(p_{i'}, y_i^*) \qquad \text{(Equation 9)}$$

B is the batch size, y_i^* is the pseudo-label from the other model, C_x is calculated from the prediction probabilities of the other model, and Du is the unlabeled set. The CPSA loss is a more straightforward case of the CPSA with confidence weighting, where the weighting C_x is always 1. Furthermore, I only use those unlabeled images which contain eosinophils for the CPSA methods, as estimated by the baseline model trained on the labeled set. Since learning the representation of eosinophils is the primary goal, cutting down on images without any eosinophils helps reduce the class imbalance. This loss is computed for each model with respect to the pseudo-label provided by the other model.

Figure 5.3 shows the training setup for CPSA.

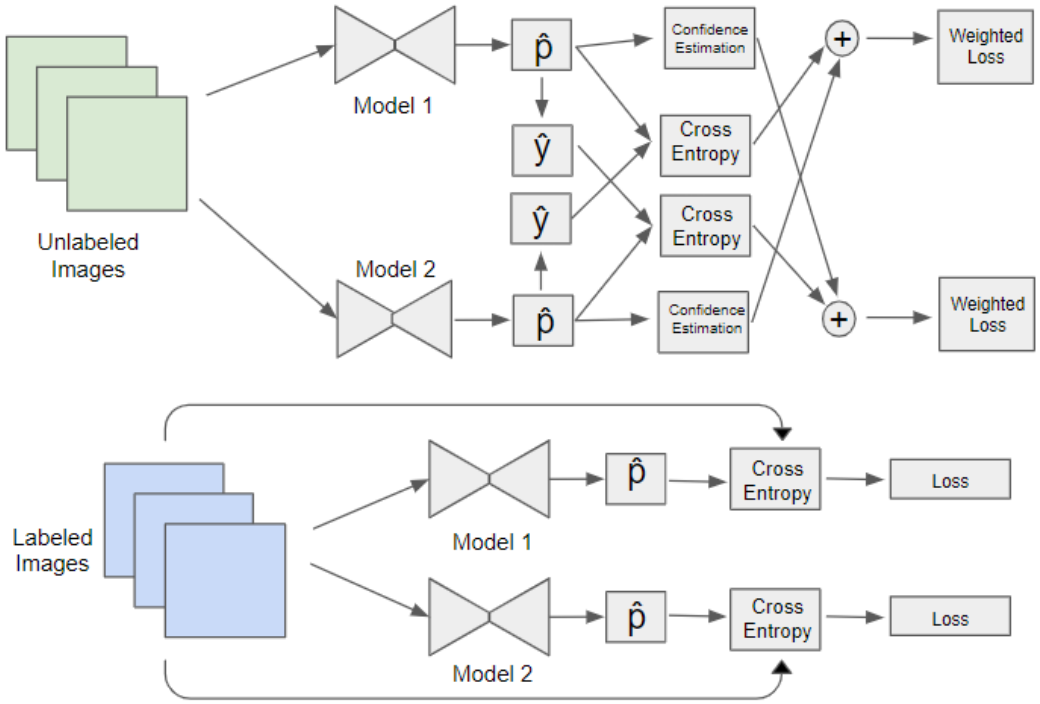

Figure 5.3: This figure shows the cross-pseudo supervision augmented with confidence estimation training.

5.5.3 Clustering Approach

I hypothesize that a lack of generalizability of the SSL methods is due to a difference in histological features between the patients' datasets. This hypothesis has grounding in the medical literature, which has observed histological features and their distributions differ by race and ethnicity (Li et al., 2002; Schwimmer et al., 2005). To examine the patients' feature distributions, I implement the following method. For each dataset, I pass the unlabeled images through a classifier pre-trained via self-supervised contrastive loss on 57 datasets, most of which are H&E stained (Ciga et al., 2022), as in Figure 5.4.

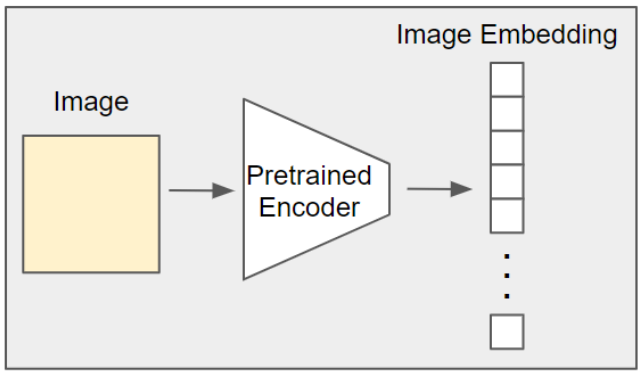

Figure 5.4: This figure shows the process of projecting an image into the embedding space.

The output is a feature vector with 512 components. The feature embeddings of the unlabeled patches are clustered via a Gaussian Mixture Model (GMM). To learn the parameters of the Gaussians, the GaussianMixture package from the scikit-learn library (Pedregosa et al., 2011) in Python is used. This uses the k-means algorithm to initialize the means of the Gaussians. Then, the expectation-maximization (EM) algorithm is used to iteratively update the parameters of the Gaussians until convergence. This is performed for 2 to 19 clusters.

To determine the optimal number of clusters amongst that range, I select the number that minimizes the Bayesian information criterion (Schwarz, 1978), given the unlabeled data and learned model parameters. Equation 10 defines the measure.

$$BIC = k * log(n) - 2log(L) \qquad \text{(Equation 10)}$$

k is the number of model parameters of the Gaussian distributions, n is the number of unlabeled images and L is the maximum likelihood of the unlabeled image feature embeddings given the learned model parameters, as calculated in the bic function of the scikit-learn package. Increasing the number of Gaussians results in a higher likelihood of the data, but it also causes an increase in the number of parameters, which is penalized. Figure 5.5 shows the clustering process and selecting the optimal number of clusters.

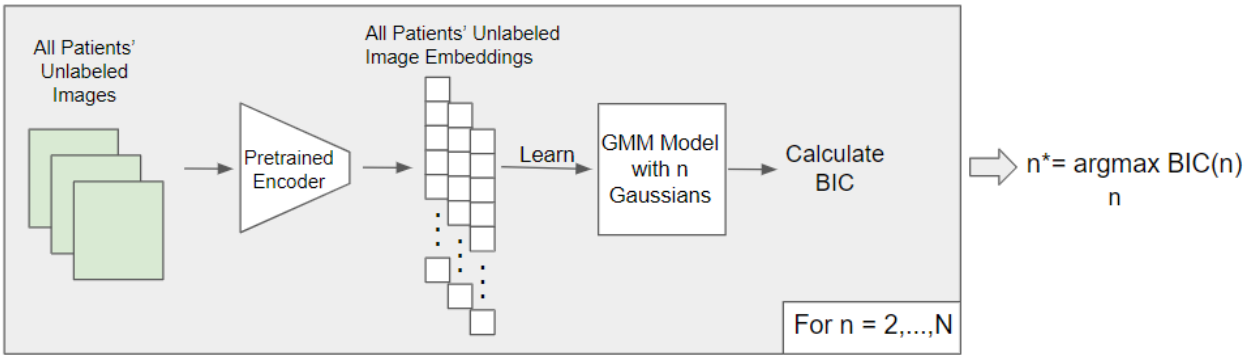

Figure 5.5: Finding the Optimal Number of Gaussian Clusters

Once the optimal number of Gaussian distributions is found, I use the learned GMM to cluster the unlabeled and labeled patches.

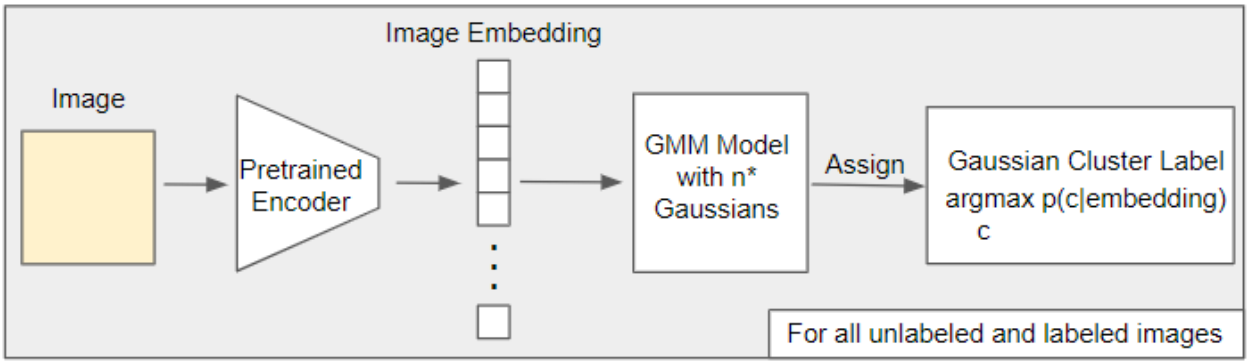

Figure 5.6: Assigning Cluster Labels using the Learned GMM

Once all the images have been assigned to a cluster, I analyze the percentages of images per patient in each cluster. Notably, the Gaussian parameters are learned from the unlabeled data only, as opposed to a mix of the labeled and unlabeled data in Kurian et al. (2023). If both labeled and unlabeled sets were used in this work, it would need to be performed for each patient, using the labeled data from the patients in the training set and the unlabeled data from the held-out patient. This would increase the complexity of the analysis since an optimal number of clusters would need to be found for each patient. Not only would the number of clusters likely be different, but also, the clusters would not be comparable across patients. Thus, standardizing the number and parameter values of the Gaussian clusters by using the unlabeled dataset makes the analysis more tractable.

5.6 Experiments

5.6.1 Dataset Quantification

From both the EoE and CD datasets, I select the top 5 patients with the most labeled data for analysis. Some patients do not have many labeled images, so drawing statistically significant conclusions from their data would be difficult. Also, the models are trained and

tested per patient for multiple iterations, so capping the limit at ten total patients makes this analysis more feasible for this study. Table 5.1 shows the number of images per patient. Eo=0 means that a supervised baseline model trained on the other patients' labeled data predicted the unlabeled image to contain no eosinophils, where Eo>0 signifies the presence of eosinophils. The patients with "E" at the beginning are EoE patients; the others are patients from the CD dataset.

Table 5.1 Number of Images per EoE and CD Patients: This table shows the number of images from each patient in the labeled set and unlabeled set, as well as the unlabeled images predicted to have and not have eosinophils by a baseline model.

Patient	Labeled	Unlabeled	Unlabeled$_{Eo>0}$	Unlabeled$_{Eo=0}$
E-17	71	486	251	235
E-105	63	909	562	347
E-139	41	1082	469	613
E-116	37	1164	503	661
E-25	33	801	737	64
RK10028	33	648	481	167
INCR0009E	15	1705	1110	595
INCT0016D	14	355	157	198
RK10349	13	909	460	449
RK10373	11	564	141	423

5.6.2 Implementation Details and Evaluation Metrics

For the supervised baseline, CPSA, and CP models, I use a U-Net++ architecture initialized with the EfficientNet-B0 encoder pre-trained on ImageNet, a learning rate of 5e-4, a batch size of 8, and binary cross-entropy loss. For each patient's baseline model, which is trained on the labeled set of the other patients, I train for 85 epochs. I train the CPSA and CPSA with confidence weighting models for 25 epochs for the CD dataset and 40 epochs for the EoE dataset. Following Peng et al. (2020), I train the CPSA models only using labeled data for a warm-up period of 10 epochs before incorporating the unlabeled data and confidence estimation. This allows the model to find a sufficient signal resulting in more informative and less noisy pseudo-labels. Furthermore, Chen et al. (2021) present no clear batch sampling strategy for CPS. Since the unlabeled set is much larger than the labeled set for all patients, combining them into a single pool from which to sample batches is not practical. Instead, I follow the sampling procedure in Yang et al. (2021), where half the batch is sampled with replacement from the labeled set and half from the unlabeled set without replacement for pseudo-labeling. This allows the labeled set to continue providing a baseline signal to anchor the learning. Additionally, the validation set consists only of labeled data and does not include any data from the test patient for the pseudo-labeling methods. I use the dice coefficient to compare the baseline supervised and semi-supervised techniques, averaged across three runs.

5.7 Results

5.7.1 Segmentation Performance

The CPSA method with confidence weighting achieved the optimal performance for five patients, three in the EoE dataset and two in the CD dataset. The CPSA without confidence weighting was optimal for three patients, 1 in the EoE dataset and 2 in the CD dataset. The supervised baseline was optimal for patient E-25 in the EoE dataset and RK10373 in the CD dataset, as shown in Table 5.2. The confidence weighting helps performance in most patients, but not all. Still, my primary focus is that no SSL method could improve the segmentation performance for the two patients mentioned, E-25 in the EoE dataset and RK10373 in the CD dataset. Furthermore, these two patients have the lowest supervised segmentation performance among patients in their respective datasets. Next, I turn to examining the clustering results.

Table 5.2: This table shows the model performance on the EoE and CD datasets via dice coefficient per patient across multiple methods.

Patient	Baseline	CPSA	CPSA + Confidence Weighting
E-17	66.3	65.7	**68.8**
E-105	76.1	76.0	**76.5**
E-139	59.0	**62.8**	57.4
E-25	**55.8**	53.7	49.4
E-116	73.5	73.1	**74.1**
RK10028	59.5	**62.1**	59.6
INCR0009E	61.7	**62.5**	61.8
INCT0016D	57.1	59.4	**59.6**
RK10349	56.7	56.6	**58.8**
RK10373	**38.5**	30.6	29.0

5.7.2 Clustering Results

The optimal number of clusters was 4 for both the EoE and CD datasets. Figures 5.7 and 5.8 show the Bayesian information criteria scores for GMMs with 2 to 19 clusters for the EoE and CD datasets, respectively, averaged over three runs.

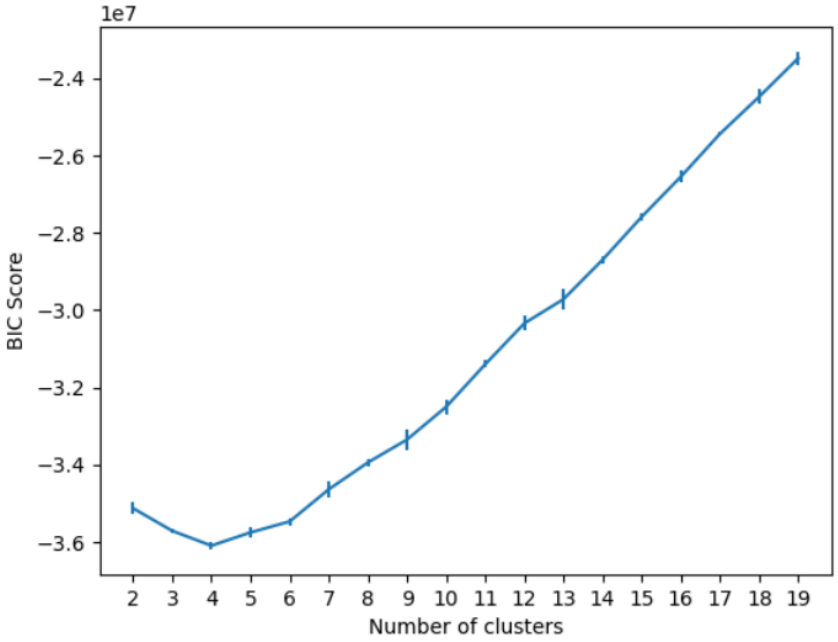

Figure 5.7: BIC Scores of GMMs for EoE Dataset

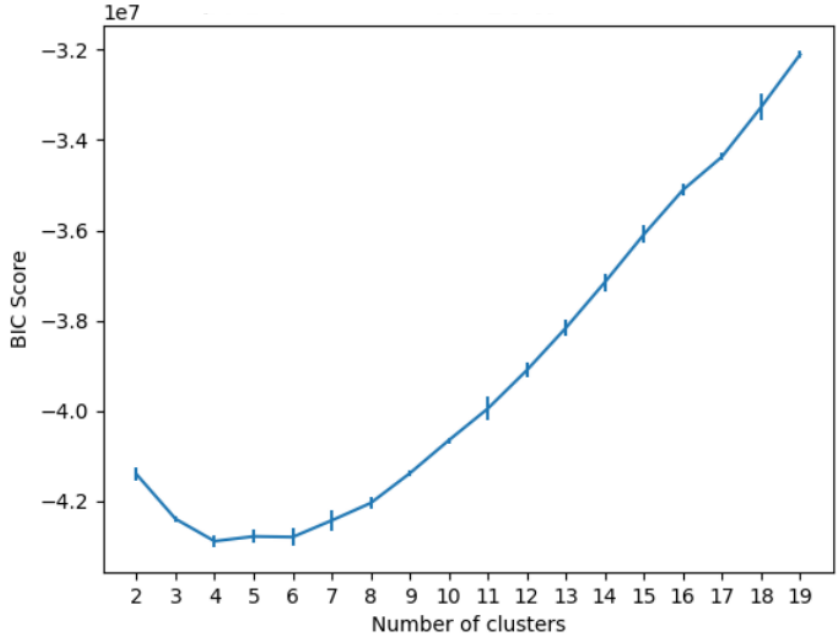

Figure 5.8: BIC Scores of GMMs for CD Dataset

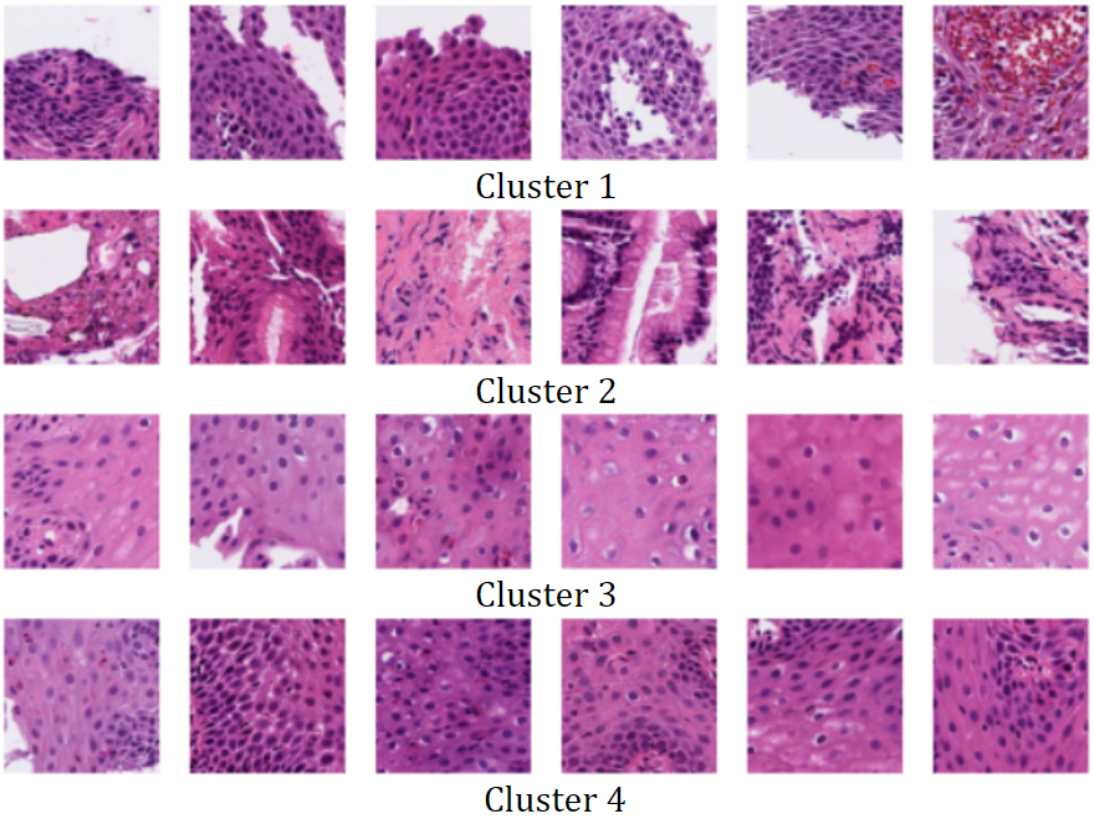

Figure 5.9: This figure shows samples from each of the clusters in the EoE dataset.

Figure 5.9 shows samples from the 4 cluster types, with six examples per row. Each row is a separate cluster. Table 5.3 shows the percentage of images belonging to each cluster in the EoE and CD datasets, as well as the percentage of total labeled eosinophils in each cluster. For the patients where the SSL methods did not outperform the supervised baseline method, E-25 and RK10373, there were clear abnormalities in the distribution of their images among the learned clusters.

Table 5.3 shows that for E-25, no images in the labeled set come from cluster 4. Furthermore, only 2% of the unlabeled images belonged to cluster 4. Figure 5.9 shows that cluster 4 contains images densely populated with nuclei cells. Table 5.3 shows that this cluster makes up 61% of the eosinophils in the other EoE patients. Cluster 4 has the majority of eosinophils for all labeled data. Also, the second largest number of images are in cluster 2 for the labeled and unlabeled set for E-25. Upon review from a doctor in my lab, cluster 2 represents tissue in the epithelial layer. This is the topmost layer and has the smallest number of eosinophils, at 6%. Thus, E-25 has an over-representation of images from cluster 2, which has the least amount of eosinophils, and an under-representation of images from cluster 4, which has the most number of eosinophils. This demonstrates a sampling bias in the data for E-25, where the biopsy was not deep enough but primarily taken from tissue with the least signal. This bias has an interpretable clinical solution of taking another biopsy that captures deeper tissue.

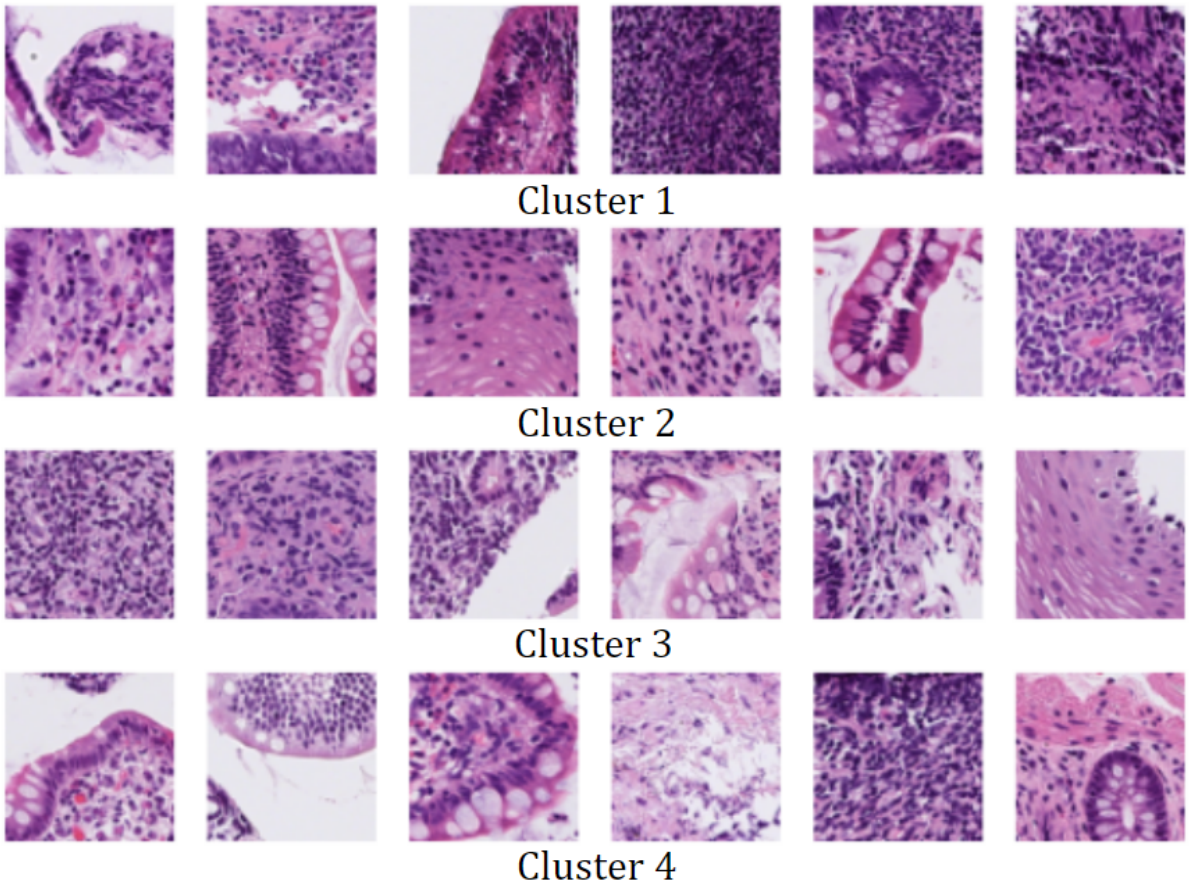

Figure 5.10: This figure shows samples from each of the clusters in the CD dataset.

For RK10373 in the CD dataset, no labeled images occur from cluster 1 in the labeled set, which has the highest amount of eosinophils. However, 29% of the images from RK10373's unlabeled set are in cluster 1. The issue here is not a sampling bias but a labeling bias. None of the images from cluster 1 were labeled for this patient. Instead, only the sparsest clusters were labeled, making the test set more difficult. The clinical solution here would be to label more data from cluster 1 for patient RK10373.

5.8 Conclusion

In this chapter, I applied SSL methods to address segmenting eosinophils without using a given patient's labeled data in training or validation. I found that for patients where the SSL methods failed to improve performance, either a sampling or labeling bias occurred. Though each bias dealt with the sparsity of eosinophils, my distributional analysis identified precisely where it occurred. The implications for the clinical solutions to address these biases are profound. Additional labeling would likely not help with the sampling bias for E-25, since my clustering analysis revealed that the biopsy did not contain a sufficient amount of deep tissue, where the eosinophils are more likely to occur. Conducting another biopsy for patient RK10373 would be unnecessary since the tissue with the most eosinophils exists but is not yet labeled. If another biopsy were taken, the

bias would not be addressed unless the images from the deeper clusters were labeled. My clustering analysis helps us understand which clinical solutions would have the greatest efficacy in increasing model performance for each patient.

Table 5.3: Clustering Unlabeled and Labeled Data on EoE and CD: This table shows the percentage of image patches belonging to each cluster for each patient. The percent of eosinophils per cluster is presented for the labeled set as well.

Patient	Unlabeled				Labeled			
Cluster	1	2	3	4	1	2	3	4
E-17	47%	39%	1%	14%	66%	0%	6%	28%
E-105	30%	22%	0%	48%	19%	0%	8%	73%
E-139	16%	5%	19%	59%	10%	7%	34%	59%
E-25	65%	33%	0%	**2%**	79%	**21%**	0%	**0%**
E-116	22%	28%	3%	47%	3%	0%	24%	73%
Percent of Eosinophils					23%	**6%**	9%	**61%**
RK10028	31%	26%	13%	31%	70%	9%	0%	21%
INCR0009E	29%	25%	5%	41%	40%	13%	0%	47%
INCT0016D	30%	30%	9%	31%	57%	7%	0%	33%
RK10349	31%	27%	9%	33%	62%	15%	0%	23%
RK10373	**29%**	26%	13%	32%	**0%**	45%	0%	55%
Percent of Eosinophils					**57%**	17%	0%	26%

Chapter 6
Conclusion

In this dissertation, I addressed optimizing segmentation performance in histopathology with limited labeled data. To address that question, I broke the problem into three subcomponents: model initialization, incorporating unlabeled data with reliable uncertainty quantifications, and identifying biases in the data. The model initialization results showed that using weights pre-trained with ImageNet was optimal when labeled histopathology was limited. This is likely due to differences in the features between the source and target histopathology datasets. These differences occur due to various forms of institutional bias, such as variance in staining techniques and magnification, and domain shifts, such as biopsies from different organs. This finding is helpful since most model architectures are pre-trained with ImageNet. Also, since no histopathological equivalent of ImageNet exists, the requisite dataset would need to be gathered, and time-consuming pre-training would need to be performed.

Next, I incorporated unlabeled data by adapting the uncertainty quantification from the classification setting. Notably, the correlation between the model's entropy estimation and the dice performance on the training dataset proved helpful in communicating the trustworthiness of the uncertainty quantification. Specifically, when the correlation was negative, the model could not be trusted to give an informative uncertainty quantification for prioritizing unlabeled images. This can be understood as a type of miscalibration of the model, exacerbated by the sparsity of the segmentation dataset.

Finally, I addressed the issue of identifying bias in the unlabeled and labeled sets, which hinder the generalization of semi-supervised segmentation models. I focused my study in the context of individual patients, replicating the real-world setting where a patient's whole-slide image may not have any labels. My clustering process used only the unlabeled set for learning the parameters, which allowed for the comparison of clusters across patients. Analyzing the clusters revealed different types of biases, namely labeling and sampling bias, which required different clinical interventions for two outlier patients from the EoE and CD datasets.

I conclude by considering the broader implications of this work. The question considered at the outset recognizes that the amount of labeled data is limited. While I worked within this constraint with semi-supervised approaches, such a constraint need not persist. Rather, a standard community dataset for histopathology images could be generated, similar to ImageNet. While privacy is often cited as a primary roadblock to this, the existing published unlabeled data could be labeled to produce a standardized set. The more difficult task, in my view, is a consistent taxonomy with ubiquitous medical relevance. A starting point could be nuclei segmentation and classification, as mentioned in this work. Once such a dataset existed, a more thorough comparison to using ImageNet for pre-training could be conducted.

Another important consideration of this work is the focus on the trustworthiness of a model's uncertainty quantification. This relates to the greater need for transparency in the model's decisions. If a model's decisions cannot be explained well, we may be unable to

detect when inherent biases in the data adversely influence those decisions. As we enter an age of increasing reliance on statistical and generative machine learning models, we must continue to develop means of establishing and verifying the trustworthiness of models.

Also, I approach bias identification on a patient-by-patient basis. This is a paradigm rarely considered in the literature. Namely, a patient's unlabeled dataset can enable tailoring an algorithm to maximize performance on that specific patient. As we move towards personalizing healthcare, our solutions should be tailored to each patient. Hopefully, this work will encourage greater consideration of each patient's feature distribution that can be appropriately and thoroughly characterized.

Bibliography

Aceves, S. S. (2011). Tissue remodeling in patients with eosinophilic esophagitis: What lies beneath the surface? *Journal of Allergy and Clinical Immunology*, 128(5), 1047-1049.

Alhmoud, T., Gremida, A., Steele, D.C., Fallahi, I., Tuqan, W., Nandy, N., Ismail, M., Altamimi, B.A., Xiong, M.J., & Kerwin, A. (2020). Outcomes of inflammatory bowel disease in patients with eosinophil-predominant colonic inflammation. *BMJ Open Gastroenterology*, 7(1), e000373.

AlZoubi, O., Tawalbeh, S. K., & Al-Smadi, M. (2020, October 16). Affect detection from Arabic tweets using ensemble and deep learning techniques. *Journal of King Saud University - Computer and Information Sciences*.

Bejnordi, B. E., Litjens, G., Timofeeva, N., Otte-Höller, I., Homeyer, A., Karssemeijer, N., & Van Der Laak, J. A. (2016). Stain specific standardization of whole-slide histopathological images. *IEEE transactions on medical imaging*, *35*(2), 404-415.

Bidgoli, A. A., Rahnamayan, S., Dehkharghanian, T., Grami, A., & Tizhoosh, H. R. (2022). Bias reduction in representation of histopathology images using deep feature selection. *Scientific Reports*, 12, 19994.

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory* (pp. 92-100). ACM.

Bucci, T. J. (2002). Basic techniques. In *Handbook of Toxicologic Pathology*, 2nd ed. (Haschek, W. M., Rousseaux, C. G., Wallig, M. A., Eds.), vol. 1, chap. 8, pp. 171-185. Academic Press, San Diego, CA.

Burkhardt, J. E., Pandher, K., Solter, P. F., Troth, S. P., Waite-Boyce, R., Zabka, T. S., & Ennulat, D. (2011). Recommendations for the evaluation of pathology data in nonclinical safety biomarker qualification studies. *Toxicol Pathol.*, 39, 1129–1137.

Calderon-Ramirez, S., Oala, L., Torrents-Barrena, J., Yang, S., Elizondo, D., Moemeni, A., Colreavy-Donnelly, S., Samek, W., Molina-Cabello, M. A., & López-Rubio, E. (2023). Dataset Similarity to Assess Semi-supervised Learning Under Distribution Mismatch Between the Labeled and Unlabeled Datasets. *IEEE Transactions on Artificial Intelligence*, 4(2), 282-291.

Calderon-Ramirez, S., Yang, S., Elizondo, D., & Moemeni, A. (2022). Dealing with distribution mismatch in semi-supervised deep learning for COVID-19 detection using chest X-ray images: A novel approach using feature densities. *Applied Soft Computing*, 123, 108983. doi: 10.1016/j.asoc.2022.108983. PMID: 35573166; PMCID: PMC9085448.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In H. Daumé III & A. Singh (Eds.), *Proceedings of the 37th International Conference on Machine Learning* (pp. 1597-1607). PMLR. (Vol. 119 of Proceedings of Machine Learning Research)

Chen, X., Yuan, Y., Zeng, G., & Wang, J. (2021). Semi-supervised semantic segmentation with cross pseudo supervision. In *CVPR*.

Ciga, O., Xu, T., & Martel, A. L. (2022). Self-supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7.

Ciompi, F., Geessink, O., Bejnordi, B. E., De Souza, G. S., Baidoshvili, A., Litjens, G., Van Ginneken, B., Nagtegaal, I., & Van Der Laak, J. (2017). The importance of stain normalization in colorectal tissue classification with convolutional networks. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)* (pp. 160-163). IEEE.

Conze, P.-H., Brochard, S., Burdin, V., Sheehan, F. T., & Pons, C. (2020). Healthy versus pathological learning transferability in shoulder muscle MRI segmentation using deep convolutional encoder-decoders. *Computerized Medical Imaging and Graphics*, *83*, 101733.

Dehkharghanian, T., Bidgoli, A. A., Riasatian, A., Mazaheri, P., Campbell, C. J. V., Pantanowitz, L., Tizhoosh, H. R., & Rahnamayan, S. (2023). Biased data, biased AI: deep networks predict the acquisition site of TCGA images. *Diagnostic pathology*, *18*(1), 67.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Li, F.-F. (2009). ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition (pp. 248-255).

Devan, K. S., Walther, P., von Einem, J., Ropinski, T., Kestler, H. A., & Read, C. (2019). Detection of herpesvirus capsids in transmission electron microscopy images using transfer learning. *Histochemistry and cell biology*, *151*, 101-114.

Everingham, M., Eslami, S. A., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2015). The Pascal visual object classes challenge: A retrospective. *International journal of computer vision*, *111*, 98-136.

Fischer, A. H., Jacobson, K. A., Rose, J., & Zeller, R. (2008). Hematoxylin and eosin staining of tissue and cell sections. *CSH protocols*, *2008*, pdb.prot4986.

Gamper, J., Koohbanani, N. A., Benet, K., Khuram, A., & Rajpoot, N. (2019). PanNuke: An Open Pan-Cancer Histology Dataset for Nuclei Instance Segmentation and Classification. In *Digital Pathology* (pp. 11-19).

Graham, S., Vu, Q. D., Raza, S. A., Azam, A., Tsang, Y. W., Kwak, J. T., & Rajpoot, N. (2019). Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, *58*, 101563.

Gonsalves, N. P., & Aceves, S. S. (2020). Diagnosis and treatment of eosinophilic esophagitis. *Journal of Allergy and Clinical Immunology*, 145(1), 1–7.

Hägele, M., Seegerer, P., Lapuschkin, S., Bockmayr, M., Samek, W., Klauschen, F., Müller, K.R., & Binder, A. (2020). Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. *Scientific reports*, *10*(1), 1-12.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

Komura, D., & Ishikawa, S. (2018). Machine Learning Methods for Histopathological Image Analysis. *Computational and Structural Biotechnology Journal*, 16, 34-42.

Kurian, N. C., S, V., Patil, A., Khade, S., & Sethi, A. (2023). Robust Semi-Supervised Learning for Histopathology Images through Self-Supervision Guided Out-of-Distribution Scoring. *arXiv preprint arXiv:2303.09930*.

Lafarge, M. W., Pluim, J. P. W., Eppenhof, K. A. J., Moeskops, P., & Veta, M. (2017). Domain-Adversarial Neural Networks to Address the Appearance Variability of Histopathology Images. In et al. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, DLMIA ML-CDS 2017, Lecture Notes in Computer Science(), vol. 10553. Springer, Cham.

Lee, D.-H. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop on Challenges in Representation Learning*.

Li, C. I., Malone, K. E., & Daling, J. R. (2002). Differences in breast cancer hormone receptor status and histology by race and ethnicity among women 50 years of age and older. *Cancer Epidemiology Biomarkers & Prevention*, 11(7), 601-607.

Lin, H., Chen, H., Dou, Q., Wang, L., Qin, J., & Heng, P. (2018). ScanNet: A Fast and Dense Scanning Framework for Metastatic Breast Cancer Detection from Whole-Slide Image. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 539-546).

Oliver, A., Odena, A., Raffel, C. A., Cubuk, E. D., & Goodfellow, I. (2018). Realistic Evaluation of Deep Semi-Supervised Learning Algorithms. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 31). Curran Associates, Inc.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

Peng, J., Estrada, G., Pedersoli, M., & Desrosiers, C. (2020). Deep co-training for semi-supervised image segmentation. *Pattern Recognition*, 107, 107269.

Qiao, S., Shen, W., Zhang, Z., Wang, B., & Yuille, A. (2018). Deep co-training for semi-supervised image recognition. *Proceedings of the European Conference on Computer Vision (ECCV)*, 135–152.

Ray, I., Raipuria, G., & Singhal, N. (2022, July). Rethinking ImageNet Pre-training for Computational Histopathology. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 3059-3062). IEEE.

Reynolds, D. A. (2009). Gaussian mixture models. In *Encyclopedia of Biometrics* (Vol. 741, pp. 659-663).

Rizve, M., Duarte, K., Rawat, Y. S., & Shah, M. (2021). In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *International Conference on Learning Representations*.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 234–241). Springer.

Rosenberg, H. F., Dyer, K. D., & Foster, P. S. (2013). Eosinophils: changing perspectives in health and disease. *Nature Reviews. Immunology*, 13(1), 9-22.

Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, 6(2), 461-464.

Schwimmer, J. B., Behling, C., Newbury, R., Deutsch, R., Nievergelt, C., Schork, N. J., & Lavine, J. E. (2005). Histopathology of pediatric nonalcoholic fatty liver disease. *Hepatology*, 42(3), 641-649. doi: 10.1002/hep.20842

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E. D., Kurakin, A., Zhang, H., & Raffel, C. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Conference on Neural Information Processing Systems*.

Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J. M., Ciompi, F., & van der Laak, J. (2019). Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis, 58*, 101544.

Xie, B., Li, S., Li, M., Liu, C. H., Huang, G., & Wang, G. (2022). Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation. *arXiv preprint arXiv:2204.08808*.

Yang, L., Zhuo, W., Qi, L., Shi, Y., & Gao, Y. (2021). St++: Make self-training work better for semi-supervised semantic segmentation. *arXiv preprint arXiv:2106.05095*.

Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., & Shinozaki, T. (2021). Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *Advances in Neural Information Processing Systems*, 34.

Zheng, X., Fu, C., Xie, H., Chen, J., Wang, X., & Sham, C.-W. (2021). Uncertainty-Aware Deep Co-training for Semi-supervised Medical Image Segmentation. arXiv preprint arXiv:2111.11629.