Domain Adaptation Evaluation for Deep Image Segmentation

Kevin J. Lin

A dissertation for the
Doctor of Philosophy degree
presented to the faculty of the
School of Data Science
University of Virginia
on 19 May 2024

Dissertation committee

Donald Brown, Advisor

Stephen Baek, Chair

Sana Syed

Afsaneh Doryab

Christopher Moskaluk

Domain Adaptation Evaluation for Deep Image Segmentation

by

Kevin Je-Kuan Lin

The views expressed in this document are those of the author and do not reflect the official policy or position of the United States Air Force, the United States Department of Defense, or the United States Government.

Domain Adaptation Evaluation for Deep Image Segmentation

Kevin Je-Kuan Lin

(ABSTRACT)

Advances in deep learning approaches for medical image segmentation show increasingly impressive results for disease detection. More specifically, these approaches have demonstrated a strong capability to detect and quantify cellular features despite significant differences in diseases, location, histopathology staining, size of the data, and imaging techniques. The goals of this dissertation are to develop a framework for medical image segmentation approaches, compare an existing domain adaptation approach with other methods, and create a new approach for future domain adaptation research. The first part of this dissertation focuses on evaluating the dice scores of the leading medical image segmentation model, U-Net, and using Monte Carlo Dropout to produce an entropy quantification metric to quantify and visualize areas where this model has difficulty in detection. Applicability of this approach is first proven with a baseline Brain Tumor Segmentation (BraTS) challenge 2021 dataset and then verified through segmentation evaluation on the private Eosinophilic Esophagitis (EoE) dataset. The created Monte Carlo Dropout U-Net maintains comparable dice scores on the EoE dataset and allows for a visualization of the entropy which highlighted detected cells and areas of interest. The second part of this dissertation focuses on extracting the entropy metric from the first part of this dissertation to separate observations into domains representing varying levels of entropy, using these domains to create a Multi-Domain Adversarial Network (MDAN)[105], and comparing this MDAN performance to that of a Denoising Diffusion Probabilistic

Models (DDPM)[36]. The motivation for the MDAN approach stems from the fact that adversarial network approaches are robust to the lack of available training data common in deep medical image segmentation. The third part of this dissertation introduces a new domain adaptation approach named Extremity-Ranked Domain Selection (ERDS) which ranks observations by their extremity and performs a full factorial experimental design to evaluate the impact of this domain choice on a MDAN dice score. An observation has high extremity if removing that observation's data in training has a large impact on the performance of the model. Observation extremity represents a new but important parameter in domain adaptation approaches for medical image segmentation. By successfully evaluating and creating domain adaptation techniques, both the medical and data science field benefit through greater understanding of how current deep medical image segmentation approaches detect diseases. Through these findings, future researchers can trust and leverage domain adaptation techniques on image segmentation applications.

Dedication

To Mom & Dad, Frank, and Grandpa.

Acknowledgments

Endless thanks to the medical researchers of the UVA Gastroenterology (GI) Data Science Laboratory for obtaining the EoE dataset through biopsies of UVA Medical Center patients. Specifically, I would like to thank Fisher Rhoads and Dr. Sana Syed for their patience and diligence in providing key medical information for this dissertation. Additionally, I am grateful for both the GI Data Science Laboratory and the UVA School of Data Science which funded international conference travel to Barcelona and Singapore respectively. I would not have been able to attend these conferences and perform at the highest level possible without all your support. Research was supported by National Institutes of Health (NIH) through the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) under award numbers K23DK117061-01A1 (Syed) and R01DK132369 (Syed).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Biomedical Image Segmentation

Deep image segmentation has been a proven approach to help pathologists understand disease progression [91]. The process starts with medical providers taking biopsies from patients to help inform a patient's condition. These biopsies are placed on slides, stained, and finally digitized, creating whole-slide-images (WSIs) which are then divided into smaller "patches." A visualization of a contemporary WSI scanner, biopsy slides, and a patch from a biopsy image is shown in Figure 1.1. Medical researchers

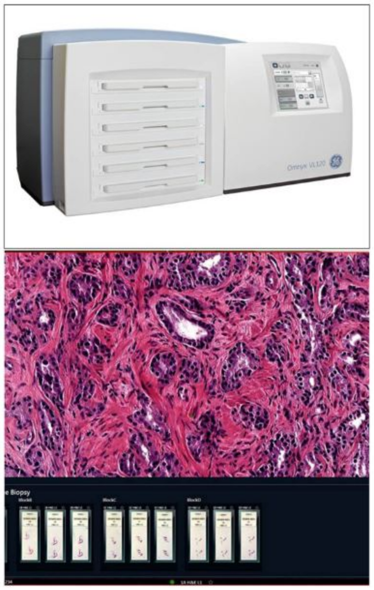


Figure 1.1: Whole Slide Imaging (WSI) Scanner, Biospy Slide, and Patch Creation[21]

then annotate these patches, “labeling” key cell types and extracting disease information. This “labeling” can be extremely costly as WSIs datasets are significantly large. For example, a typical WSI may contain 100,000x100,000 pixels and patients may have multiple WSIs for a biopsy [88]. Therefore, researchers only label small portions of the entire WSI per patient. As these “labeled” cell types typically have similar characteristics, there is significant potential for deep learning, more specifically image segmentation, to assist pathologists in labeling and extracting disease information while reducing the cost of analyzing these WSIs. Current approaches in medical image segmentation have demonstrated the performance of various architectures including ResNet [35], DenseNet[75], and PCANet [10]. In 2015, Ronnenburger created the U-Net Convolutional Neural Network (CNN) specifically for biomedical image segmentation and cell-tracking [71]. Modern U-Net architectures including Vanilla U-Net, Residual U-Net, R2U-Net, and Attention U-Net have produced impressive results [2].

Like other machine learning models, U-Nets are highly sensitive to different domains. More explicitly, the performance of a U-Net trained from a particular source domain can drop (or increase) significantly if transferred to a different target domain. Additionally, similarities in source domain do not necessarily mean similarities in U-Net performance in the target domain. In the world of medical image segmentation, providing medical researchers complete information about a disease phenotype must be the priority given the direct impact a diagnosis has on patients. Increasing prediction results can be a difficult process with common approaches such as increasing model size and tuning model parameters risking significant issues in overfitting and cost. One of the most significant limitations of segmentation methods is that the predicted label output shows only distinct classes. In the binary class case, all pixels are either classified as 0 or 1. However, this result is produced through a simple threshold value over predicted probabilities. For example, a threshold value of 0.5 would mean any pixel with a predicted probability of less than 0.5 would be classified as 0 and any pixel with a predicted probability of 0.5 or greater would be classifed as 1. This creates a deceivingly simple output map when the model is actually outputting probabilities that widely vary over the values from 0 to 1. A representation of this entropy would give valuable information into how deep segmentation models detect cells or areas of interest. On the medical side, this entropy would give medical researchers insight into deep phenotyping of cells in different observations, informing fields such as histology and morphology.

Continuing with this effort of providing more information to data scientists and medical researchers, leveraging the proper source domain for a given target domain can provide opportunities to consider observation-level characteristics. There are many different approaches for domain adaptation but it is currently unclear which methods are most appropriate for deep learning computer vision problems[92][39][15]. Most researchers perform domain adaptation evaluations for convolutional neural network (CNN)-based approaches on very small images[73]. This is likely due to the computational rigor of evaluating larger images. Therefore, it is still largely unknown what optimization methods are feasible and most efficient for segmentations involving large, high-resolution images. Since segmenting large medical images typically requires dividing the image into smaller patches for each observation, domain adaptation can be a possible approach to determine which observations contribute most to a model's ability to identify cells or areas of interest. The model can be further evaluated through an entropy metric to analyze how intentionally sub-setting observations can drive segmentation results.

1.2 Problem Statement

The goal of this dissertation is to evaluate the effectiveness and feasibility of various domain adaptation methods on the image segmentation application. For clarification, this work will be testing various combinations of source and target domain choices and sizes, focusing on observation-level dice score evaluation and entropy quantification. This work will not be focused on evaluating all current approaches in domain adaptations. Currently, there is a lack of interpretability of medical deep learning models where even though a predicted label is created, little information is given as to what is actually happening in the model when creating this prediction. This research will directly address this gap through an entropy visualization finding areas in images where deep learning models are likely to make incorrect predictions. Furthermore, there is currently an overreliance on training models that treat all observations as equal. Recent research on major datasets such as a Lizard, PanNuke, and CoNSep, focuses more on detecting and identifying cells or areas of interest regardless of observation-level analysis. This observation-level analysis includes consideration of individual patient characteristics. This focus on patients is justified through the National Institutes of Health (NIH)'s priority of "patient-centered care [that] focuses on the individual's particular health care needs" [70] and former President Barack Obama's $215 Precision Medicine Initiative (PMI) in 2015. The PMI "takes into account individual differences in people's genes, environments, and lifestyles"[64] verifying that the highest levels of the United States Government support observation-level approaches. Grouping all patients in datasets and creating a blanket model over all patients' training data should not reflect actual medical practices and should not be what data science researchers pursue.

The findings of this research can be extended to any deep computer vision task that requires large and high-resolution image inputs. The evaluation of these methods will assess trade-offs between model performance, computation time, and memory resources required. Additionally, finding an appropriate method for domain adaptation can minimize the need for more training data to improve performance. Particularly in the medical imaging field, training data can be costly to obtain due to privacy concerns.

To demonstrate that the most efficient domain adaptation techniques can compare well across wide-ranging datasets, these models will be evaluated on various datasets from current research. The first dataset will be brain MRI images taken from the Brain Tumor Segmentation (BraTS) challenge 2021 dataset which consists of images that are 240x240 large and are separated into 28,075 images for training and 3,043 images for testing. This dataset will be treated as the baseline dataset to train models since numerous researchers have used this dataset to assess segmentation performance [101][66][6][40][20]. The second dataset consists of annotations of eosinophil white blood cells in biopsy images for patients diagnosed with Eosinophilic Esophagitis (EoE). Detecting and quantifying eosinophils is essential to diagnosing and managing the EoE disease. The dataset is obtained from the Gastroenterology Data Science Lab from UVA Hospital patient data. Each image is 512x512x3 large and there are 514 images/masks in the dataset spanning 30 UVA Medical Center patients.

1.3 Dissertation Overview

The following is the overview of this dissertation. The second chapter is the literature review that presents image segmentation approaches, evaluation metrics, and domain adaptation techniques. The third chapter details segmentation approaches and entropy visualizations for two different biomedical datasets. The fourth chapter explains the methodology of the experimental setup and introduction of the Extremity-Ranked Domain Selection (ERDS) method. The fifth chapter provides the results of the ERDS and lists the performance improvement from using a observation-level metric to create source and target domains. Finally, the sixth chapter provides conclusions and explores future work for consideration.

This dissertation contributed to many fields including data science, image segmentation, machine learning, deep learning, computer vision, artificial intelligence, medical image analysis, domain adaptation, and more. The following is a list of some of the major contributions:

- Demonstrated through experimentation that Bayesian optimization, specifically through Monte-Carlo Dropout, can produce entropy visualizations per patch that provide insight into the abilities of deep segmentation approaches and into the deep phenotyping present in medical data observations.
- Evaluated through experimentation that Multi-Domain methods have comparable performance to Diffusion based methods while providing more control over model training. This control allows observation-level analysis in determining the effect each observation has on deep segmentation model results.
- Created a new metric for evaluating deep learning approaches by dropping subsets of training data and evaluating the effects of these omissions. This metric is a direct representation of a deep learning model's reliance on a subset of a dataset and can provide valuable information about what model performance will be when certain features of a dataset are not present.

- Showed the relationship of domain size, domain choice, and threshold value in domain adaptation. Full factorial designs encompass each pairing of these parameters driving efforts to improve future domain adaptation techniques. Furthermore, segmentation results vary between different source/target domain pairings. Intentionally setting source and target domains based off of the impact of each observation improves deep segmentation results.

- Produced optimal domain adaptation parameters that verified the importance of the new patient observation extremity (POE) metric on deep learning models. Setting these parameters significantly increased a multidomain adaptation method's dice score, demonstrating that the model is able to better identify areas of interest in an input image. Accompanying observation-level entropy quantification results provide information into how each observation contributes to model results.

- Used a traditional statistical approach through a full factorial design to evaluate a multi-domain adaptation approach created by the new extremity metric. Bonding a traditional approach with a modern one and achieving strong results indicates that future work should consider using traditional statistical and mathematical methods to support current approaches.

- Improved the detection performance of two different biomedical image segmentation projects that all have critical clinical importance.

There is a lack of research on evaluating domain adaptation techniques on medical images particularly for multi domain methods[1][80]. The motivation behind most domain adaptation research has been to address the cost needed to collect and annotate large scale training data and minimize the possible shift between training and test samples. This has lead to the creation of single-source, single-target approaches focusing on knowledge transfer from a labeled source domain to an unlabeled target domain and exploring domain-invariant structure and representations[46]. However, the labeled data for many of these domain adaptation techniques may come from multiple domains with different distributions. As a result, naive application of the single-source-single-target domain adaptation algorithms may lead to suboptimal solutions. Such problem calls for an efficient technique for multiple source domain adaptation. Additionally, there have also been problems in medical image segmentation research with not having sufficient labeled data in a domain to produce accurate results[48]. Multidomain adaptation approaches have the potential to use other source domains in addition to domains with insufficient labeled data to bridge this gap. Additionally, multidomain adaptation approaches can allow deep segmentation models to focus on differences in observations. Most medical segmentation approaches treat all observations (which contain data from patients) equally which creates a disconnect with the medical field that focuses treating each observation as unique. Medical segmentation approaches typically focus only on detecting and identifying areas of interest independent of observation characteristics. Another way to view this difference is that medical care is given and adjusted based on patient-specific information. Care that works for one patient is not guaranteed to work on another patient. Medical deep learning approaches must also then consider approaches that can focus on observation-level analysis and must also minimize grouping large amounts of observations as equal. Other approaches to improve performance on deep learning computer vision such as hyperparameter optimization still rely on large amounts of labeled data for training.

The increased knowledge from this research will be useful for many computer vision applications. Multi domain adaptation methods can produce insight into which domains contribute most in segmentation approaches. Additionally, there is currently a lack of research listing which source/target domains provide the best performance[8]. Testing different combinations of source/target domains will give insight into which pairings are most effective and potentially give more information regarding domain adaptation models. Future work in medical imaging would benefit from knowing which source/target interactions perform best, eliminating the need for traditional trial and error approaches. If successful, this research has the potential to minimize and potentially eliminate the requirement for large amounts of labeled data to diagnose diseases while cementing trust in the biomedical community for deep learning models.

Chapter 2

Review of Literature

Research related to this dissertation falls into three major categories: U-Net image segmentation models, Monte Carlo Dropout, and domain adaptation. This chapter focuses on the literature that falls into these categories, connects these approaches to real-world datasets, and explores the various approaches researchers are using for analysis. Emphasis will be on exploration of deep learning architectures, mathematical formulation, and comparison of performance metrics.

2.1 U-Net Image Segmentation

The U-Net is convolutional network architecture for fast and precise segmentation of images [71]. It has been shown to outperform what was previously considered the best method (a sliding-window convolutional network) on the ISBI challenge for segmentation of neuronal structures in electron microscopic stacks. Typically in experiments, the UNet model utilizes 23 convolutional layers with batch normalization. In both the encoder and decoder steps, ReLU is used as the activation function, and for the final layer, sigmoid is used as the activation function. Loss is computed using binary cross entropy and ADAM is used as the optimizer. Early stopping and data augmentation are implemented to prevent overfitting. The concept relies on the baseline U-Net architecture shown in Figure 2.1, using the encoder to obtain and normalize the transformation of the input volume, using a Leaky ReLU activation function at each layer. At the bottleneck of this architecture, the volume will be in the size of 2×2×2 which represents the reduction of dimensionality prior to using a sigmoid activation function for segmentation. The decoder then up-samples this transformed 2×2×2 volume to reconstruct the image with this segmentation.

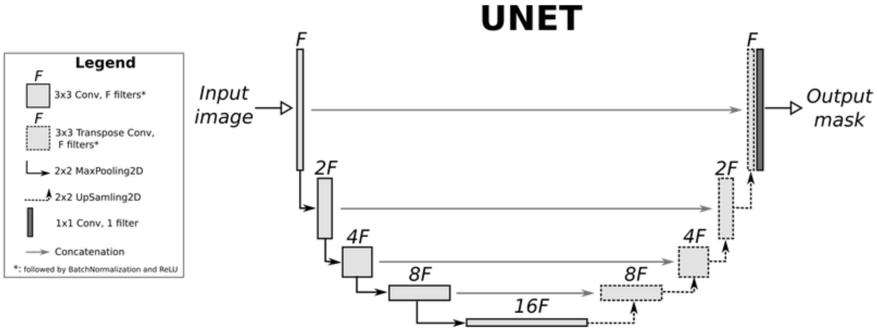


Figure 2.1: UNet Architecture: An input image is passed through an encoder and decoder in order to create an output mask.[41]

The loss function used for training U-Net models is typically an Intersection-over-Union (IoU) metric. These techniques tend to outperform cross entropy when the segmentation is sparse or a small fraction of the total image. One of the most popular IoU techniques is the Dice similarity coefficient[19]. The dice coefficient is the industry and academic standard for evaluating medical image segmentation and classification results. Evaluation is given through a scale of 0 to 1. Given two images X and Y, a zero dice coefficient indicates that there is no similarity between two images while a one dice coefficient indicates that the images are exactly the same down to the pixel. The equation for the metric is given by

$$\text{DCS} = \frac{2|X \cap Y|}{|X| + |Y|} \tag{2.1}$$

Most state-of-the-art dice scores on medical images range from 0.3 to 0.7 [95][92].

Much like a regression metric, a dice score of 0 or 1 are usually causes for concern since no two images from real-world data are truly the same. Although classification accuracy is often used as an evaluation metric, it is important to note that the dice score may be low even if the overall image training and validation classification accuracy are high on a given dataset due to significant class imbalances.

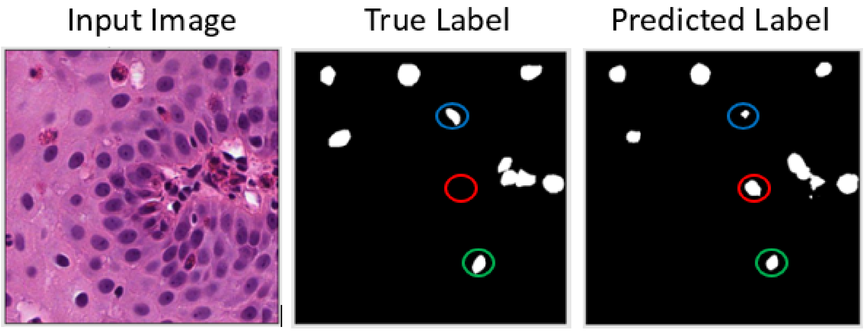


Figure 2.2: Segmentation Evaluation of an UNet approach. The green circle shows an area where the UNet was able to identify the correct location, size, and shape of the cell. The red circle shows an area where the UNet incorrectly identified a cell. The blue circle shows an area where the UNet correctly identified the correction location of the cell but incorrectly identified the size and shape of the cell.

A visual representation of this evaluation is shown in Figure 2.2. Most importantly from this visualization is the fact that the UNet makes mistakes and sometimes siginficant ones. The red circle where the UNet mistakenly detected a cell of interest can be a serious issue when medical providers are relying on correct information to assist in disease diagnosis. This research will directly address minimization efforts for incorrect detection through an entropy metric in the next section.

2.2 Monte Carlo Dropout

With the majority of the dataset stemming from WSIs and other significantly large sources, scalability and efficiency play a vital role in performance. Due to the significantly large dataset, variational approximation, specifically minimizing the Kullback-Leibler divergence (KL divergence), is needed in order to approximate the distribution of the affected cells. The KL divergence is a measure of how different two probability distributions are. KL divergence is given by

$$\mathbf{KL}(q(\mathbf{Z}||p(\mathbf{Z}|D))) = -(E_q(\log p(D,\mathbf{Z})) - E_q(\log q(\mathbf{Z}))) + \log p(D) \tag{2.2}$$

Equation 2.2 illustrates that the KL divergence is just the expected log likelihood ratio. Since maximizing the evidence lower bound (ELBO) is largely impractical, Monte-Carlo approximations are commonly used. Work from Gal and Ghahramani in 2016 [28], suggest that a Monte Carlo Dropout U-Net is equivalent to the deep Gaussian process used in Bayesian Neural Networks. Essentially, the KL divergence can be minimized using approximation through Monte Carlo integration to get an unbiased estimate. Minimizing the KL divergence between the approximate posterior $q(w)$ and the posterior of the full deep Gaussian Process $p(w|X,Y)$ is given by the objective function:

$$-\int q(w)\log p(Y|X,w)dw + KL(q(w)||p(w)) \tag{2.3}$$

The first and second term can be represented by a sum and approximated by Monte Carlo integration. For the Monte Carlo Dropout U-Net, dropout is applied before every weight where dropout is defined as switching off neurons at each training step.

In Bayesian neural networks, each weight is represented by a probability distribution which is assumed Gaussian instead of just a number. The learning aspect corresponds to Bayesian inference which uses MC Sampling. Entropy is then calculated for every pixel using cross-entropy over two classes of “background” ($C = 0$) (e.g. not pixels containing the cells of interest) and “foreground” ($C = 1$) (e.g. pixels containing the cells of interest) :

$$U = -(p_{C=0} \cdot \ln(p_{C=0}) + p_{C=1} \cdot \ln(p_{C=1})) \tag{2.4}$$

As an additional performance metric, model entropy will verify consistency in the data. This is given by

$$E_{p(z|D)} H[p(y|z,x)] = -\int p(z|D) \left(\sum_{y \in Y} p(y|z,x) \log p(y|z,x) dw \right) \tag{2.5}$$

Through literature review, deep learning conferences and researchers refer to this metric as “uncertainty” [44][18] but closer observation into the underlying mathematical formula yields that this “uncertainty” is merely the same equation as entropy. In fact, DeVries and Taylor defined model uncertainty as the following:

Model uncertainty z is estimated by calculating the entropy of the averaged probability vector across the class dimension:

$$z = -\sum_{c=1}^{C} p_c \log p_c \text{ where } p \text{ is the probability vector and } c \text{ are the classes} \tag{2.6}$$

This “uncertainty” metric has proven to be extremely important in deep segmentation. Kendall states that “The model’s uncertainty is an effective measure of confidence in prediction"[44] and DeVries follows with "Uncertainty estimates are capable of detecting when a neural network is likely to make an incorrect prediction"[18]. In deep medical segmentation, this measure of how likely a model will make an incorrect prediction is a significant motivator as medical researchers are relying on correct information to assist in deep phenotyping efforts. Some visualizations to assist this discussion are shown in Figures 2.3 and 2.4. In both figures, the uncertainty visualizations provide insight into what areas the deep segmentation models struggle to predict.

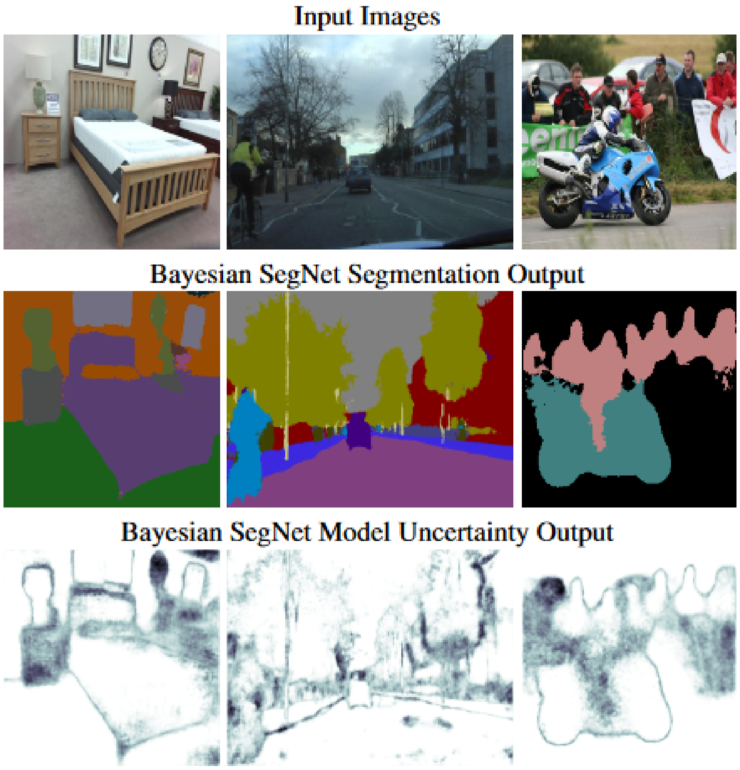

Input Images
Bayesian SegNet Segmentation Output
Bayesian SegNet Model Uncertainty Output

Figure 2.3: Uncertainty Visualizations from a Bayesian SegNet on various datasets[44]. High uncertainties are observed at class boundaries and when objects are visually difficult to identify or appear visually ambiguous.

Although deep learning researchers have referred to this metric as "uncertainty" in multiple articles. A more proper term would be entropy and this was clarified through

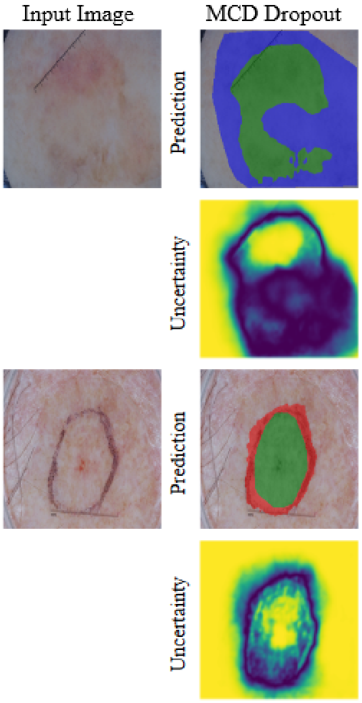


Figure 2.4: Uncertainty Visualizations from a Monte Carlo UNet on a skin lesion dataset[18]. The high uncertainties observed at class boundaries have caused significant errors. The red boundary surrounding the skin lesion represents a false positive in classifying the pixel as a skin lesion when there is not one present in that location.

an exchange between Claude Shannon, the "father of information theory" and the famous mathematician, John von Neumann, in 1961. Shannon states:

> "My greatest concern was what to call it. I thought of calling it 'information,' but the word was overly used, so I decided to call it 'uncertainty.' When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, 'You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one knows what entropy really is, so in a debate you will always have the advantage.' "[85]

In the spirit of one of the greatest mathematicians of all time, the word “entropy” will be now be used instead of “uncertainty” to represent this metric with the only exceptions being published works already using the word “uncertainty.”

2.3 Entropy in Medical Image Segmentation

The concept of entropy has been well established in the fields of thermodynamics, biology, and information theory. Specifically in the field of information theory, entropy is usually defined as the ”Shannon entropy” in respect for Claude Shannon[47]. The core concept of information theory is that information transmitted by a message can be measured. Closely associated with this concept of information is probability. Taking the example of measured messages, defined a message as m each message is then associated with a corresponding probability p_m, or the probability that the message is extracted from a set of messages. Shannon entropy is then defined as the amount of information which is missing before reception[78]. Formally, let $H(p_m)$ be the entropy of a set of messages m, then

$$H(p_m) = -\sum p_m \log p_m \tag{2.7}$$

This formula is exactly the same as the one used in Equation 2.6, linking the mathematical formulation behind machine learning and information theory. Most significantly, this Shannon entropy is well regarded as a powerful tool for optimizing the amount of information involved in the transmission of messages or the storing of data. As an expansion for this work, Shannon and Weaver further introduced optimum coding, which depends on the probabilities and on a possible noise that may destroy part of the messages[78]. For convenience, Shannon entropy will now be referred to as merely entropy in this work. Entropy can also be a measure of the "health" of a system through measuring the knowledge observers have of the system[60]. For example, entropy is low in the case of skewed probability distributions because one event dominates. Conversely, entropy is high in the case of balanced probability distributions where no event dominates another since the observer has little knowledge of what the next measure will be. Medical image segmentation problems tend to have serious class imbalances with the number of positive samples being much smaller than the number of negative samples[50]. Directly, this means that more of the pixels in the image are associated with the background instead of being associated with the cells being detected. This research directly addresses this issue through the choice of the Dice Similarity Coefficient evaluation metric given in Equation 2.1 which is a more robust metric in imbalanced problems than other evaluation metrics such as accuracy.

Arguably the most familiar entropy concept in machine learning is cross-entropy. Cross-entropy loss is one of the most common methods of evaluating a model's performance[103][50][12]. Cross entropy is typically used to adjust model weights during training and closely related to the KL divergence shown in Equation 2.2. However, KL divergence calculates the relative entropy between two probability distributions while the cross entropy calculates the total entropy between the distributions[67]. As stated prior, the KL divergence is used in this research's Monte Carlo Dropout approach where the goal is minimization in order to approximate the distribution of the affected cells. Also, as stated before, the cross entropy is not used in this research because there are significant issues with using this metric on imbalanced problems[50]. In general, the medical datasets used in this research contain far more background pixels (e.g. not pixels containing cells of interest) than foreground pixels (e.g. containing cells of interest). To close, this research will use entropy for medical image segmentation approaches in two ways: 1) as evaluation metric to inform locations where a model is likely to make an error and 2) KL divergence (relative entropy) to approximate the distribution of the affected cells. Therefore, these entropy concepts will be crucial for both segmentation datasets in Chapter 3 and for setting up

2.4 Domain Adaptation

Deep learning[52] has greatly pushed forward the development of artificial intelligence and machine learning[45] . As one of the most popular deep learning models, the convolutional neural network (CNN) has demonstrated its superiority over conventional human-engineered imaging features[22][56][108][97][96]. Trained with large-scale labeled data in a full supervision manner, the CNN has made breakthroughs in computer vision [102][54][11][86] and medical image analysis[1][93][42][106]. It has been revealed that the CNN is able to learn generic low-level features (e.g., textures and edges) that can be transferable to different image analysis tasks [31]. These advancements in deep learning extend to computer vision methods where domain adaptation has shown impressive results. For supervised domain adaptation, the standard approach is to transfer models learned on a source domain onto the target domain and then fine tune the model to try to match the target domain. Some of the approaches included in literature include PCANet[10], ResNet[75], and U-Net[71]. The advantages of these methods is that having labelled data gives the models a lot more power in prediction[65]. However, the labels also can potentially create irregularities in the dataset [31]. One solution here is to partition each class within the image dataset into a set amount of subsets and then assign new labels to the new set. Furthermore, collecting and annotating such large-scale training data is both prohibitively expensive and time-consuming [105]. To solve these limitations, different labeled datasets can be combined to build a larger one, or synthetic training data can be generated with explicit yet inexpensive annotations. However, due to the possible shift between training and test samples, learning algorithms based on these cheaper datasets still suffer from high generalization error[53]. Domain adaptation (DA) focuses on such problems by establishing knowledge transfer from a labeled source domain to an unlabeled target domain, and by exploring domain-invariant structures and representations to bridge the gap. Figure 2.5 shows an example of how transfer learning, which domain adaptation is a subset of, differs from traditional machine learning approaches. Whereas most traditional machine learning approaches create only task-specific models, transfer learning strives to extract information about a given (source) task and match it to a different (target) task[13][30][69][33]. The motivation behind this approach is that if a model can capture the similarities between the source and target tasks, new models do not have to be created when new tasks appear. These similarities are called "domain invariant features." In the case of deep learning where training a new model means significant resource requirements, the motivation behind a transfer learning or domain adaptation method becomes immediately apparent.

Furthermore, domain adaptation techniques are typically divided into two different fields: divergence based domain adaptation methods and adversarial based domain adaptation methods. Divergence based domain adaptation methods assumes that fine-tuning the deep network model with labeled or unlabeled target data can diminish the shift between the two domains. This fine tuning is done through minimizing some divergence-based criterion (e.g. KL divergence) between the source and target distribution. Ideally, this would extract domain invariant features. The advantage

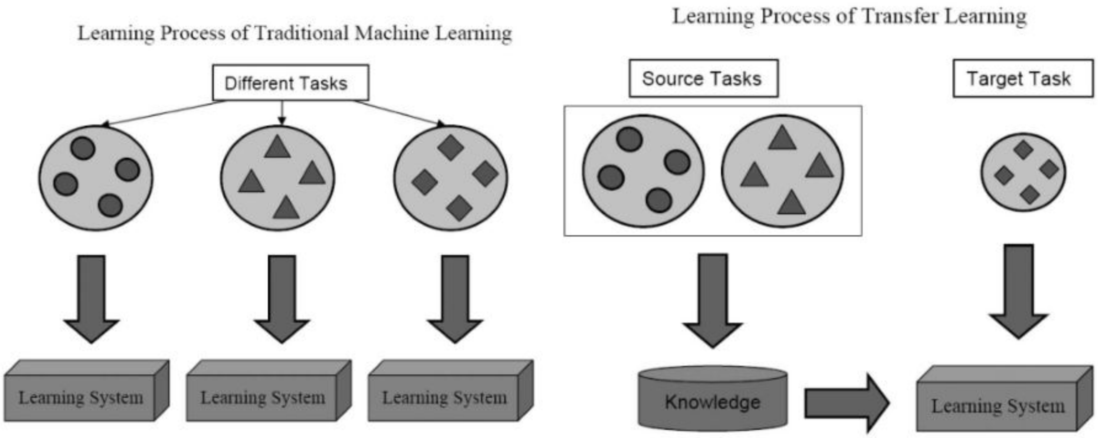


Figure 2.5: Transfer Learning Comparison: Traditional Machine Learning relies on a “one task, one model” whereas transfer learning tries to use models for multiple tasks by minimizing the distance between a source and target task

here is that if the divergence can be learned based on a given dataset, domain adaptation is expected to perform better than other methods. However, the trade-off is that the divergences are usually non-parametric and not specific to the dataset [89]. This can lead to the approach taking a significant amount of time to train if the source and target domains are significantly different from each other.

Adversarial based domain adaptation shares many similarities to the Generative Adversarial Network (GAN) with a generator and discriminator. Applying this concept to domain adaptation, the generator is just a feature extractor and the discriminator networks attempt to distinguish between source and target domain features. A gradient reversal layer (GRL) is commonly added to the network in order to evaluate both gradients in one standard backpropagation step. The main advantage of the using adversarial based domain adaptation is that it is comparatively simple for implementation. The GRL is easily added to standard Deep Learning models without further modifications. However, the issue is that the adversarial assumption limits all experiments to only two domains. Extending this work to other domains will require parameter adjustments. Additionally, training can be unstable due to the adversary training scheme. A solution here would be to use momentum as optimizer [89]. GAN methods also address one of the most common issues with domain adaptation methods which is domain shift[95]. An example of an issue with domain shift would be a drop in model performance on a model trained on a certain source domain when applied to a different target domain.

Most research focuses on single source and single target domain adaptation for segmentation[87][46] but little work has been done on multiple source and multiple domain methods and even less work has been done in determine ideal target domains given one or more source domains[80][59][46]. One approach that has shown promising results uses a Multisource Domain Adversarial Network (MDAN) that optimizes task-adaptive generalization bounds [105]. This approach combines the strengths of both divergence-based and adversarial-based domain methods by defining worst and average case classification bounds and using adversarial learning to extract features with an architecture shown in Figure 2.6. Zhao et al proceeded to test this network on three different datasets: Amazon Reviews, MNIST Digits Datasets, and WebCamT Vehicle Counting Dataset. In nearly cases, the MDAN outperformed comparable Domain-Adversarial Neural Networks (DANNs) shown in Figure 2.7.

Although the approach has shown promising results, the MDAN and other multi-domain methods have not been used on medical imaging datasets. Testing if this improvement holds in the case of medical imaging using datasets such as EoE biopsy slides and Brain Tumor MRI Images can give insight into how multi domain methods perform in relation to more standard approaches such as UNets. Additionally, this new approach can help guide future work for using multi domain methods in medical imaging. However, Zhao et al. did show one case under the “best-Single DANN”

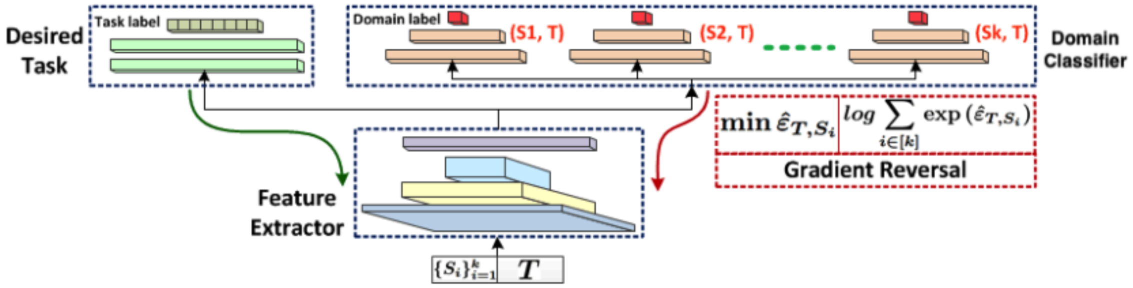


Figure 2.6: Multi-Domain Adversarial Network (MDAN) Architecture: Gradient reversal is separated into two types of approaches. Hard version: uses a source that achieves the minimum domain classification error for backpropagation. Smooth version: all domain classification risks are combined and backpropagated adaptively

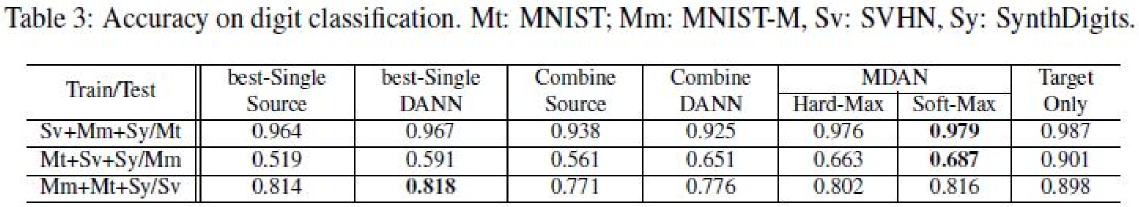


Table 3: Accuracy on digit classification. Mt: MNIST; Mm: MNIST-M, Sv: SVHN, Sy: SynthDigits.

Train/Test	best-Single Source	best-Single DANN	Combine Source	Combine DANN	MDAN		Target Only
					Hard-Max	Soft-Max	
Sv+Mm+Sy/Mt	0.964	0.967	0.938	0.925	0.976	**0.979**	0.987
Mt+Sv+Sy/Mm	0.519	0.591	0.561	0.651	0.663	**0.687**	0.901
Mm+Mt+Sy/Sv	0.814	**0.818**	0.771	0.776	0.802	0.816	0.898

Figure 2.7: MDAN Digits Datasets Results: MDAN outperforms the best Domain-Adversarial Neural Networks (DANNs) on four different digits datasets (Mt: MNIST, Mm: MNIST-M, Sv: SVHN, Sy: SynthDigits)

where the DANNs did out perform the MDAN slightly[104]. This indicates for future work that DANNs should be included in experiments to ensure that any MDANs are actually outperforming standard DANNs.

For a one-source one-target approach, define X the set of images in the dataset and Y as the set of possible labels. This research only focuses on binary labels so $Y = [0, 1]$. Define the labeling function as $f : X \rightarrow [0, 1]$. Matching images in X to the label in Y, define D_S as the source domain and D_T as the target domain. The deep learning model is then trained on the labeled data from the source domain, D_S and tested on the target domain D_T. The goal of the deep learning model is to keep a low target error. Directly, this means minimizing the probability that the model misclassifies the pixels in the image[29]. This is represented by Ben-David in 2010 in Equation 2.8[7]

$$\varepsilon_T(h) = \mathbb{E}_{\mathbb{X}\sim\mathbb{D}_\mathbb{S}}[|h(x) \neq f(x)] \tag{2.8}$$

However, this dissertation focuses on multiple source and target domains. Expanding Equation 2.8 to multiple domains yields, $\{D_{S_i}\}_{i=1}^{k}$ as the source domain and D_T as the target domain. The multidomain error is shown in Equation 2.9 and was created by Zhao et al in 2017 [105]. Breaking down the two terms in Equation 2.9, the first term asks for informative feature representations for labelling while the second term captures the domain invariant feature representations.

$$\max_{i\in[k]} (\widehat{\varepsilon}_{S_i}(h) - min_{h'\in H\Delta H}(\widehat{\varepsilon}_{T,S_i}(h')) \tag{2.9}$$

One of the core requirements for domain adaptation is that the source and target domain are different in order to properly evaluate deep learning approaches. This work will focus on source and target domains that are very similar and come from the same dataset. This means that the feature spaces between the source and target domains will have significant overlap. However, the goal of this dissertation is to assist deep phenotyping on one specific condition through analysis of different observations. Therefore, the work is focused on how factoring in certain observation characteristics that vary significantly in the dataset can actually improve a domain adaptation model's performance. In Chapter 4 of this dissertation, the observation characteristics are significantly different between both sets and individual observations and this actually motivates a domain adaptation technique in order to determine what is ultimately driving the model behavior. In Chapter 5, the domain adaptation is executed and verifies the importance of this observation-level approach. Overall, this dissertation focuses on only one medical condition in Chapters 4 and 5 in order to properly evaluate the observation characteristics and factor impacts on a domain adaptation model. This work produces optimal factor values that clearly show the observation-level analysis having a positive impact in producing effective segmentation.

The motivation for using a domain adaptation method centers on modern approaches that have shown GAN approaches performing extremely effectively for domain adaptation methods[107][99][49]. Additionally, GAN approaches can address domain shift by using a generator to project features to an image space and a discriminator operates on this projected space[89][74]. Chapter 4 of this dissertation explores a multidomain GAN approach combining the strengths of GANs on minimizing domain shift with the observation-level capability of multidomain approaches. Strong performance in Chapter 4 of this dissertation motivates using a multi domain GAN approach as an factor evaluation metric.

Chapter 3

Segmentation Datasets

This chapter illustrates two different medical imaging datasets used for segmentation. In this work, one public dataset is treated as the baseline, providing context, and one private dataset is used to introduce a new approach, providing progress in the field of deep learning. For both datasets, the technical and medical research impact are explored and presented.

3.1 Brain Tumor MRI Images

Medical image segmentation of brain tumors is one of the most challenging medical image analysis tasks due to its variable shape and appearance in multi-modal magnetic resonance imaging [101][20][81][66][6]. Manual segmentation of brain tumors requires a great deal of medical expertise, which is time-consuming and also prone to human error. On the other hand, recent improvements in Machine Learning (ML), specifically in Deep Learning (DL), help in identifying, classifying, and measuring patterns in medical images, which include image segmentation [55]. With the improvements convolutional neural networks (CNNs), many CNN models has been able to approach the human level performance in plethora of applications such as image classification or microscope image segmentation[100].

However, training an appropriate deep learning model requires not only high a quality dataset but also well designed model to learn from the data. Since UNet[71] is the most popular architecture for brain tumor segmentation, Futrega et al.[25] started their experiment with several different UNet-like architectures. They found the UNet architecture achieves the best result for the BraTS21 dataset, and they further optimized this architecture.

Training data for brain tumor detection was obtained through the Brain Tumor Segmentation (BraTS) challenge 2021 datasets. The BraTS'21 challenge bases multi-institutional pre-operative baseline multi-parametric magnetic resonance imaging (mpMRI) scans, and focuses on the evaluation of the most advanced methods for the segmentation of intrinsically heterogeneous brain glioblastoma sub-regions in mpMRI scans. Furthermore, the BraTS'21 challenge also focuses on the evaluation of classification methods to predict the MGMT promoter methylation status[6].

The BraTS 2021 data of 2,000 cases (8,000 mpMRI scans) represent a superset of the BraTS 2020 data of 660 cases (2640 mpMRI scans). Ample multi-institutional routine clinically-acquired multi-parametric MRI (mpMRI) scans of glioma, with pathologically confirmed diagnosis and available MGMT promoter methylation status (for the glioblastoma cases with such associated data), are used as the training, validation, and testing data for this year's BraTS challenge. Specifically, the datasets used in this year's challenge have been updated, since BraTS'20, with many more routine clinically-acquired mpMRI scans. Ground truth annotations of the tumor sub-regions are created and approved by expert neuroradiologists for every subject included in the training, validation, and testing datasets to quantitatively evaluate the predicted tumor segmentations. In this work, the BraTS 2021 dataset serves as a baseline dataset for segmentation as it is publicly available as a segmentation challenge dataset[5]. Therefore, performance will be compared for models trained on this dataset against future datasets.

All neural networks require adequate amounts of training data[38][51]. The greater the amount of the data, the better the classification or segmentation algorithm performs. This means that there may be limitations in access to systems with the capabilities to process all this data. For this work, there are 28,075 brain images for training and 3,043 brain images for testing. Total sizes for the two datasets amount to 670MB. In order to address the potentially large sizes of the segmentation models, the University of Virginia (UVA) Rivanna high performance computer will be used for training and testing data.

To prevent overfitting, the following image augmentation methods will be used: Biased crop (to randomly crop part of the dimensions), Zoom (to sampling from the picture and zoom the sample with cubic interpolation and mark with the most adjacent interpolation), Flips (voxels are been rotated along the axis x,y,z), Gaussian Noise (sampling each voxel with Gaussian noise and add them to the input data), Gaussian Blurring (applied on the input volume), Brightness (the input voxels are multiplied with a random value), and Contrast (the input voxels are multiplied with a random value and cropped to its original size).

After this, for each training sample, three of the MRI modalities are combined together as one multi-channel image. In order have the image size fit in the model, images are resized to to 128x128x128 and normalized. This normalization is consistent with all other approaches to the Brain Tumor Segementation (BraTS) 2021 challenge as images must be resized in order to be properly used as training data in deep learning models. To resolve any resulting issues with normalized values approaching zero, one extra voxel channel has been created, which employs the one hot encoding to distinguish the foreground and background, and this mask was added as one extra channel for each image. Training will be for 100 epochs, yielding a runtime of approximately 30 seconds per epoch.

For evaluation, loss is calculated through combination of the Binary Cross Entropy Loss and Sørensen-Dice Loss given by the Dice Coefficient.

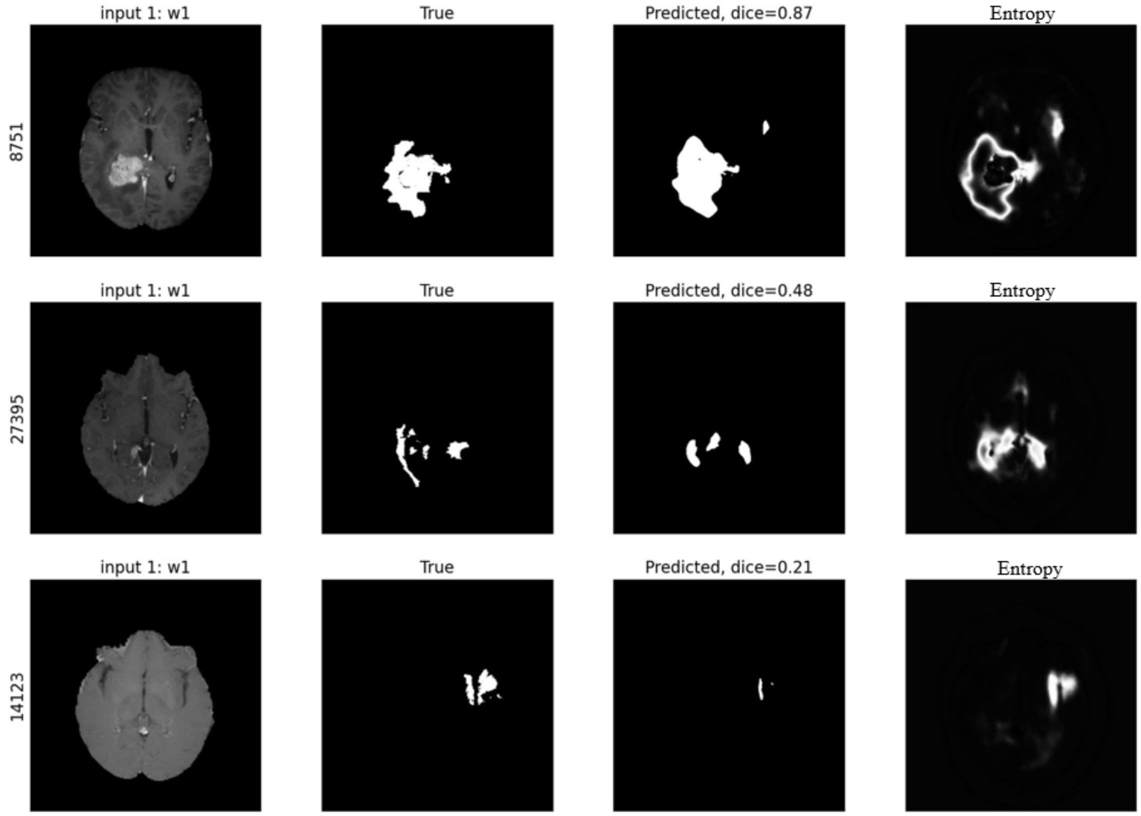


Figure 3.1: Results from Brain Segmentation: Verifies UNet Approach for Segmentation and Illustrates an Entropy Visualization

For the testing output shown in Figure 3.1, there are some cases where the model struggles on the segmentation tasks. Even in the first row of images with the highest dice score shown, there appears to be a "ghost" segmentation floating to the upper right of the correct segmentation. This segmentation may not seem significant in a general model evaluation since the overall dice score remains a respectable 0.87 but incorrect or misleading segmentations can severely impact medical staff in interpretation. In the case of brain tumor detection where results will almost certainly change a patient's life, any deep learning model must be careful to list limitations and explore segmentation results.

3.2 Eosinophilic Esophagitis Biopsy Images[58]

Eosinophilic esophagitis (EoE) is an inflammatory disease of the esophagus characterized by the prevalence of a type of white blood cell (eosinophil). Approximately 0.5-1.0 in 1,000 people have EoE and it can be seen in 2-7% of patients that undergo endoscopies [16]. Although the cause of EoE remains unclear, pathologists believe EoE to be triggered by a patient's diet. Furthermore, EoE is only increasing in prevalence [9] leading to an increased load on pathologists. Patients with EoE typically present with swallowing difficulties, food impaction, and chest pain [72][23]. A diagram illustrating the mechanism of EoE is shown in Figure 3.2.

To diagnose EoE, patients must undergo an endoscopy where eosinophils biopsy tissue samples are then evaluated for concentration of eosinophils. Pathologists diagnose the patient with EoE if at least one High-Power Field (HPF; 400× magnification adjustment) within a patient's tissue biopsy slide contains 15 or more eosinophils [24]. In order to assist in counting these eosionphils, this biopsy sample is placed on a slide and observed through a Whole Slide Image (WSI). These images can be of significant size with sizes of 100,000 x 100,000 pixels not uncommon[88]. In these WSIs, medical providers are looking for specific cells of interest which may be indicators that provide information about a patient's condition. In this work, EoE data gathered from the Gastroenterology Data Science Laboratory at the University of Virginia Medical Center spans the training dataset. The hematoxylin and eosin

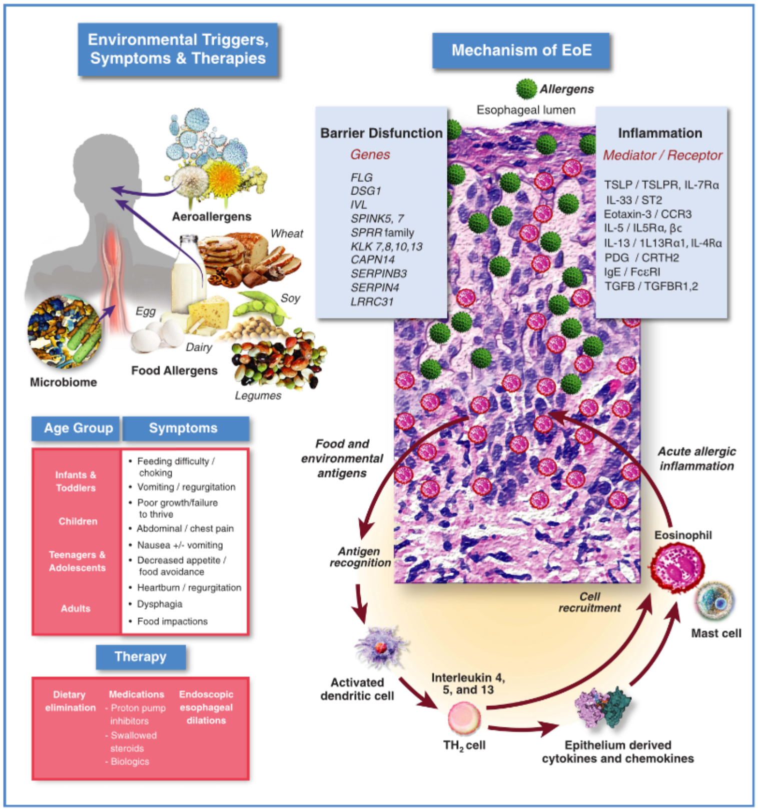


Figure 3.2: Eosinophilic esophagitis: Clinical and Pathophysiologic Overview[23]

stained (H&E) biospy images in the dataset are taken from 30 patients who have all been diagnosed with EoE by pathologists. The overall summary of the patient characteristics is given in Table 3.1. Immediately apparent is that the EoE patient data taken from UVA Medical Center seems to have observations from non-Hispanic white patients almost universally. Therefore, this research has serious limitations for applications to the wider community as underrepresented groups have zero (or nearly zero) incorporation in this dataset. Additionally, approximately a third of all EoE patients were already diagnosed with EoE before receiving the biopsy present in this dataset. This means that some patients may have been receiving treatment when their EoE data was extracted.

A sample image is given in Figure 3.3. Each image is 512x512x3 large and there

Characteristic	**UVA (n=30)**
Male (%)	18 (60)
Age in years, median (IQR)	26 (36.5-12.5)
Race	
- White, n (%)	29 (97%)
- African American, n (%)	1 (3%)
Ethnicity	
- Non-Hispanic, n (%)	29 (97%)
- Hispanic, n (%)	1 (3%)
Prior EoE Diagnosis, n (%)	9 (30%)
Treatment at Biopsy	
- Elemental Formula, n (%)	1 (3%)
- Elimination Diet, n (%)	5 (17%)
- Nasal Steroid, n (%)	0 (0%)
- PPI, n (%)	12 (40%)
- Swallowed Steroid, n (%)	3 (10%)
BMI in kg/m2, mean (SD)	27 (11)

Table 3.1: EoE Data Patient Characteristics: Almost all EoE data is taken from non-Hispanic white patients. One African American patient is represented and zero other ethnicities are present. Caution should be taken before using this research on wider samples with other ethnicities since they are not represented in the training dataset.

are 514 images/masks in the dataset spanning 30 UVA Medical Center patients. To preserve the color information for future works, the three channels [r,g,b] will be maintained for the image. However, all masks will be imported as grayscale. All data used in this work is completely de-identified and follows the United States patient privacy laws including the Health Insurance Portability and Accountability Act of 1996 (HIPAA) and the University of Virginia Medical Center's own commitment to patient privacy. Data is used for academic use only.

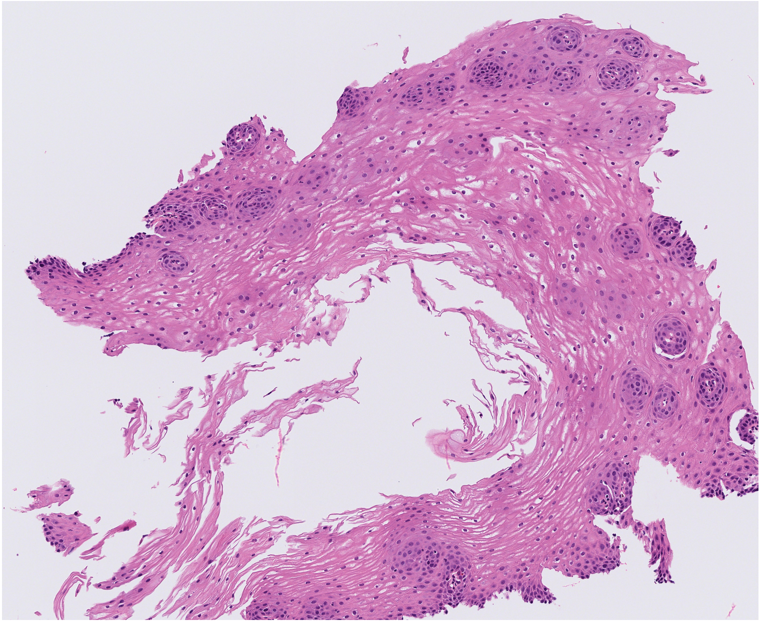


Figure 3.3: Example Image Data from Gastroenterology Data Science Lab: This biopsy image is divided into a significant amount of patches in order to assist with disease diagnosis. Pathologists then analyze these patches to assess cells of importance and gain information about a patient's condition.

The UNet is convolutional network architecture for fast and precise segmentation of images [71][51][2]. It has been shown to outperform what was previously considered the best method (a sliding-window convolutional network) on the ISBI challenge for segmentation of neuronal structures in electron microscopic stacks. To execute the experiment, the UNet model utilized 23 convolutional layers with batch normalization.

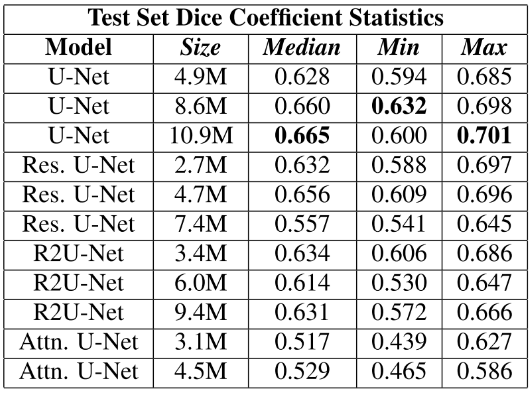
In both the encoder and decoder steps, ReLU was the activation function, and for the final layer, sigmoid was the activation function. Loss was computed using binary cross entropy and ADAM was used as the optimizer. Early stopping and data augmentation were implemented to prevent overfitting. The four models used are MCD UNet, UNet, DenseNet, and ResNet50. All training was done on 4 NVIDIA A100 GPUs with 300GB of RAM in TensorFlow/Keras 2.7. Each model run was ran for 100 epochs with a learning rate of 0.001.

At first, the encoder is used to obtain and normalize the transformation of the input volume, using a Leaky ReLU activation function at each layer. At the bottleneck of this architecture, the volume will be in the size of 2×2×2 which represents the reduction of dimensionality prior to using a sigmoid activation function for segmentation. The decoder then up-samples this transformed 2×2×2 volume to reconstruct the image with this segmentation.

To improve upon Adorno, et. al. [2], training data is augmented through flipping and rotation for each image. This means that for all input images, there exists a flipped and rotated image as well in the training data, effectively tripling the input dataset. Data augmentation here makes the model generalize better due to the larger amount of training data. For reference, the results from Adorno, et. al. [2] are shown in Figure 3.4. The size field in Figure 3.4 refers to the total number of parameters of each model. For comparison, the UNet results are shown in Table 3.2.

Model	***Size***	***Median***	***Min***	***Max***
MCD UNet	494K	0.591	0.470	0.650
UNet	494K	0.598	0.442	0.657
DenseNet	7.2M	0.592	0.369	0.645
ResNet	24M	0.612	0.505	0.651

Table 3.2: Test Results Dice Score: MCD UNet has 20x less parameters than Adorno et al's best performing UNet.



Test Set Dice Coefficient Statistics				
Model	***Size***	***Median***	***Min***	***Max***
U-Net	4.9M	0.628	0.594	0.685
U-Net	8.6M	0.660	**0.632**	0.698
U-Net	10.9M	**0.665**	0.600	**0.701**
Res. U-Net	2.7M	0.632	0.588	0.697
Res. U-Net	4.7M	0.656	0.609	0.696
Res. U-Net	7.4M	0.557	0.541	0.645
R2U-Net	3.4M	0.634	0.606	0.686
R2U-Net	6.0M	0.614	0.530	0.647
R2U-Net	9.4M	0.631	0.572	0.666
Attn. U-Net	3.1M	0.517	0.439	0.627
Attn. U-Net	4.5M	0.529	0.465	0.586

Figure 3.4: Adorno et.al. Dice Values [2]: Adorno's results illustrate that UNet approaches have strong dice scores but tend to have a significant number of parameters needed to train a model.

In comparison with literature, this approach has only ≈ 494,000 parameters while the smallest model in Adorno et al. had 3.1 million parameters and the largest had 10.9 million parameters. Thus, the work is at least one order of magnitude less in size than Adorno et al. which means the model is less complex. Given the same UNet approach, it is tempting to assume, holding all other factors such as GPU availability constant, this approach runs faster and more efficiently. However, due to the fact that the dropout layers in the MCD UNet require a non-trivial amount of time for inference, the approach will not have a full order of magnitude of increase for speed. For context, the MCD UNet took approximately 10mins to run 100 epochs on this 494,000 parameter model. Despite the significant difference is model size, the performance is comparable to Adorno et. al with dice scores around 0.591, matching their Residual UNet and R2UNet results while outperforming their Attention UNet results. This results follows current results in published literature with Gadosey et al using a stripped down UNet and producing competitive results in 2020 [26]. Additionally, these smaller model sizes can make deep learning more accessible for researchers with low computation budgets.

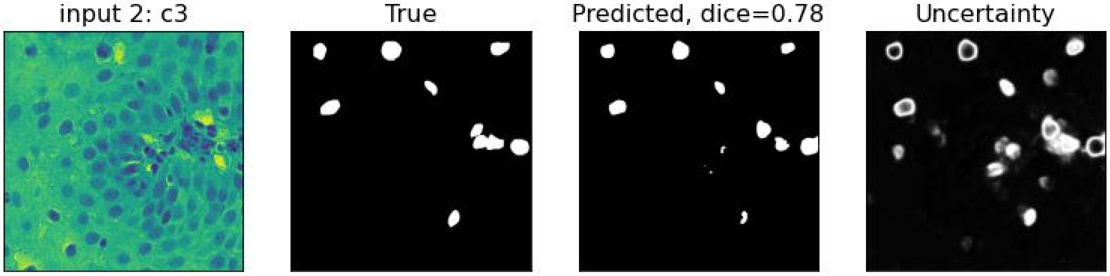


Figure 3.5: MCD UNet Results Visualization: Output comparing the true, predicted, and uncertainty values for an image patch. The uncertainty visualization demonstrates that the model has high uncertainty at the boundary of the eosinophils and may detect eosinophils even when they are not present.

Arguably, the strongest result of this approach is the visualization possible due to the quantification of the model's uncertainty is shown in Figure 3.5. Please note that this value is, in fact, the model entropy but the term “uncertainty” will be used in this section given that this paper was published using the word “uncertainty” to describe this metric. In this example image, white indicates high amounts of uncertainty while black indicates low amounts of uncertainty. Additionally, the pixels are nearly white around the borders of the eosinophils essentially “highlighting” them in the output image. The high uncertainty on boundary pixels around each eosinophil indicate that until the model sees more of the eosinophil, the model hesitates to classify the pixels as an eosinophil. This difficult are is verified by observing that once the model passes the boundary pixels of each eosinophil, the uncertainty drops significantly and the interior of the eosinophil is nearly black. This dramatic shift in values from the outside of the eosinophil which is black, to the border of the eosinophil which is white, to the interior of the eosinophil which is black again creates these “rings” of white in the resulting image. In areas where multiple eosinophils are clustered, the model seems to struggle to differentiate between the eosinophils leading to a “cloud” of high uncertainty around the area. However, this output provides valuable information to pathologists as the highlighting the general area of interest and reducing the visual load compared to the original input on the far left. These rings of white for actual eosinophils in comparison with blurred “ghost” eosinophils provides an important distinction for how deep learning models operate. Medical research has shown that deep learning models may be falsely detecting eosinophils when these cells being their “degranulation” process[14][15]. Eosinophils degranulate when they die which creates challenges for both medical researchers and deep learning models when attempting to detect and segment these cells. Particularly in the case of EoE where the diagnostic criteria is 15 or more eosinophils present in a WSI, these degranulated eosinophils can significantly impact diagnosis for medical providers and impact model training for deep learning researchers. In order to check for this possible issue, this uncertainty visualization was sent to the medical researchers at the University of Virginia Gastroenterology Data Science Laboratory who verified that while the MCD UNet uncertainty visualization seems to identify red blood cells occasionally, the approach is not significantly identifying degranulated eosinophils during model validation. Additionally, further work into the MCD UNet falsely identifying red blood cells yielded inconsistent results that had a similar impact as other noise in the dataset.

For analysis, Dice scores of all models are plotted in Figure 3.6. The boxplots overlapped, which indicates the results from each neural net are not statistically significantly better than others. This is important because uncertainty visualizations can be obtained from the MCD UNet [18][44], as shown in Figure 3.5, whereas UNet and ResNet50 do not allow for this. Monte Carlo Dropout allows a visualization of the uncertainty through the dropout layers since it randomly turns off neurons during training, which adds this stochastic element.

As the novel piece in this work, the model’s uncertainty is compared first through

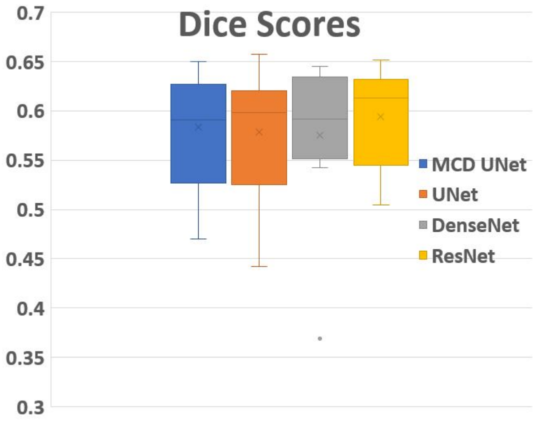


Figure 3.6: Boxplots of Model Dice Scores: All UNet approaches have similar dice scores indicating that the significantly smaller MCD UNet should be considered in future works where a vanilla UNet, DenseNet, or ResNet may be used.

an overview of the raw data and through a boxplot for visualization. Furthermore, the uncertainty of almost all models is comparatively similar with only the DenseNet having an outlier uncertainty value of 0.05 as shown in Figure 3.3. Considering the significant difference in the order of magnitude between the sizes of the models, the relative similarity in the uncertainty values demonstrate that at least at a sufficient size, model uncertainty stays consistent regardless of the size of the model chosen.

Model	***Size***	***Median***	***Min***	***Max***
MCD UNet	494K	0.007	0.004	0.016
UNet	494K	0.007	0.005	0.013
DenseNet	7.2M	0.009	0.006	0.05
ResNet	24M	0.008	0.005	0.01

Table 3.3: Model uncertainty: All UNet models have similar uncertainty results indicating that the smaller MCD UNet maintains comparative dice scores without sacrificing consistency.

As stewards of patient data, researchers in medical image analysis must ensure all models perform efficiently and appropriately. This work addresses both concerns demonstrating MCD UNet's comparable dice scores with fewer parameters than the current models and introducing model uncertainty as an evaluation metric. All mod-

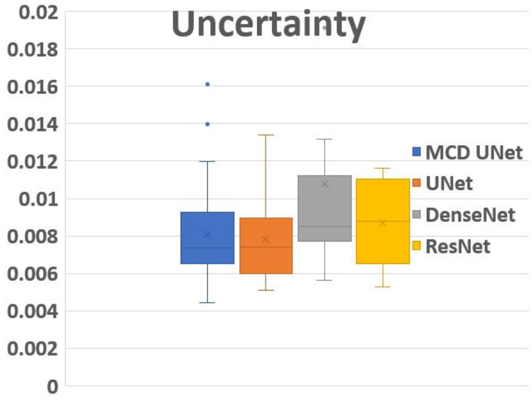


Figure 3.7: Boxplots of Model Uncertainty: Visualization of the uncertainty illustrates the similarity of UNet approaches with the exception of a few outliers.

els in this work had comparable values in uncertainty indicating that at least after a model reaches a certain size, the model uncertainty will stay constant. The model's uncertainty is then represented through a visualization which highlighted the eosinophils in the resulting image. Scaling this approach with uncertainty to full size biopsy images can help pathologists quickly identify eosinophils while also reducing the mental load. Compared to a screen full of cells and color, the black and white "rings" circling the eosinophils can at least narrow down eosinophil locations while having the potential to count all eosinophils and output a mask showing their exact locations. One of the most difficult parts of this work was working with limited observations which is a common challenge in the field of medical image analysis. A potential improvement to this work would be to incorporate few-shot learning. While the dataset exists in a high-dimensional space, the work is limited to the number of samples available due to the cumbersome nature of acquiring annotated histology images. Few-shot learning has significantly improved classification accuracy in medical imaging datasets [48] and likely would produce competitive results.

Chapter 4

Methodology

This chapter explains the approach on how domain adaptation experiments were designed and executed. The experimental setup and scenarios are presented first. The motivation for using domain adaptation techniques follows. Next, other competitive approaches towards domain adaptation techniques are discussed and evaluated. Finally, the tested evaluation metrics are listed and explained.

4.1 Experimental Setup

Prior to starting any work, appropriate work to prevent data leakage was performed to minimize overfitting. In this case, data leakage would involve allowing some of the test data to be learn off of information that would not typically be present during prediction. A common example of this would be having the true label of a model as a characteristic to be trained upon[81]. In this case, performance would be extremely high because the model already knows what to look for. However, this does not produce a useful model because, during prediction, the model will not know what the true labels look like. In fact, Singh in 2022 writes ”When data leakage occurs, it usually leads to overly optimistic outcomes during the model building phase, followed by the unpleasant surprise of poor results after the prediction model is implemented and tested on new data.” Singh continues by saying that the leakage might lead to

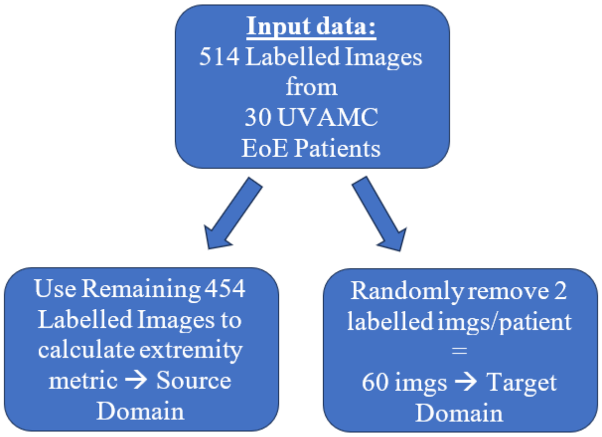

Input data:
514 Labelled Images from 30 UVAMC EoE Patients
Use Remaining 454 Labelled Images to calculate extremity metric → Source Domain
Randomly remove 2 labelled imgs/patient = 60 imgs → Target Domain

Figure 4.1: The 514 labelled images input dataset spanning 30 patients was divided into two groups in order to prevent data leakage prior to model testing.

suboptimal models being produced [81]. One way to combat leakage is by ensuring that none of the testing data will be present any model training. To that end, two labelled images are randomly selected per observation and stored separately. It is important to note that although only two labelled images have been taken from all observations, some observations may only have 10 patches whereas others may have significantly more patches. For clarification, the number of patches per patient still remains the same since the observation-level input data will not change regardless of domain selection (e.g. Observation E-123 still started with 10 patches for analysis, even though two the patches are taken for the target domain). A limitation here is that there is not a significant amount of data so storing some of the data away can be risky when deep learning models are so data-dependent when creating useful results. These 60 labelled images will eventually constitute the target domain in the domain adaptation method and this division is fully visualized through Figure 4.1. The 454 labelled images will be used to calculate extremity and entropy metrics. The extremity metric will allow separation of observations into partitions that will be this research's the source domains.

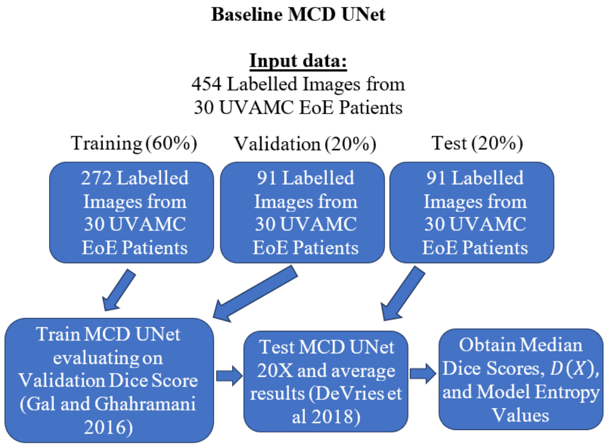

Baseline MCD UNet
Input data:
454 Labelled Images from
30 UVAMC EoE Patients
Training (60%)
Validation (20%)
Test (20%)
272 Labelled Images from 30 UVAMC EoE Patients
91 Labelled Images from 30 UVAMC EoE Patients
91 Labelled Images from 30 UVAMC EoE Patients
Train MCD UNet evaluating on Validation Dice Score (Gal and Ghahramani 2016)
Test MCD UNet 20X and average results (DeVries et al 2018)
Obtain Median Dice Scores, $D(X)$, and Model Entropy Values

Figure 4.2: The remaining images are used to evaluate each observation's extremity value which will allow control over domain choice. Training, Validation, and Test split was done 60% training, 20% Validation, and 20% test

Next, a Baseline MCD UNet is created in order to observe the dice score with all observations present in the dataset. Training, Validation, and Test split was done 60% training, 20% Validation, and 20% test. This is visualized in Figure 4.2.

Finally, the MCD UNet's dice score is evaluated through a Leave-One Out Approach where patients are one-by-one removed in training but are present in testing. Figure 4.3which shows an example of this process with observation E-123.

With the Baseline and Leave-One Out MCD UNet defined, an extremity metric can be calculated. Chapter 5 goes over the formal definition of the extremity metric and allows the creation of observation-level domains.

Specifically, in domain adaptation, this research will be focused on a multidomain approach focused on reducing the generalization error shown in Equation 2.9 back in Chapter 2. For convenience, this equation is restated below

$$\max_{i\in[k]} (\widehat{\varepsilon}_{S_i}(h) - min_{h'\in H\Delta H}(\widehat{\varepsilon}_{T,S_i}(h'))) \tag{4.1}$$

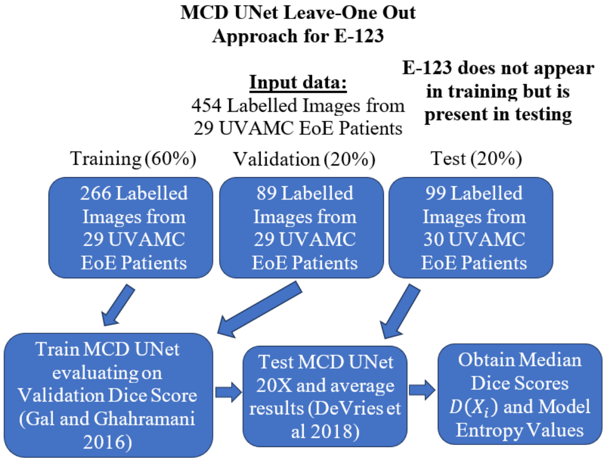

MCD UNet Leave-One Out Approach for E-123
Input data:
454 Labelled Images from 29 UVAMC EoE Patients
E-123 does not appear in training but is present in testing
Training (60%)
Validation (20%)
Test (20%)
266 Labelled Images from 29 UVAMC EoE Patients
89 Labelled Images from 29 UVAMC EoE Patients
99 Labelled Images from 30 UVAMC EoE Patients
Train MCD UNet evaluating on Validation Dice Score (Gal and Ghahramani 2016)
Test MCD UNet 20X and average results (DeVries et al 2018)
Obtain Median Dice Scores $D(X_i)$ and Model Entropy Values

Figure 4.3: MCD UNet Leave-One Out Approach for Analyzing Observation-Level Data: For this example, data for observation E-123 is removed in training but present in testing.

The generalization error shown in Equation 4.1demonstrates the domain adaptation model will be attempting to find informative feature representations for labeling while capturing the domain invariant feature representations [105]. The first term is the empirical source error while the second term represents the Once again, the source domains are created by placing observations in groups of high and low extremities and the target domain is the 60 reserved images taken from each observation in the input dataset. At this point, these 60 images have not been used for training, testing, or validation in any model for this research which means active efforts were made to prevent data leakage. Most importantly, in the field of domain adaptation, the source and target domains are clearly separated. The resulting model will then be attempting to generalize to an unseen target domain which will provide more information about how medical conditions appear in deep segmentation models.

One critique of this observation-level Leave-One-Out approach is that it seems similar to a different concept called K-fold cross-validation[82]. Leave-One-Out methods have historically been associated heavily with K-fold cross validation methods with major Python packages such as sklearn clearly stating that a ”LeaveOneOut()” cross-validator function is equivalent to ”KFold(n_splits=n)”[76]. However, there is literature of Leave-One-Out approaches that are not cross validation. ICLR 2024 had a paper titled ”Leave-One-Out Distinguishability in Machine Learning” by researchers in the National University of Singapore (Source: https://arxiv.org/pdf/2309.17310) that also measured the influence of training data points in machine learning, a concept that is core to this dissertation[98]. In this paper, the words ”K-fold”, ”CV” or ”cross-validation” do not appear and the focus is on measuring observation-level data influence and addressing issues with data leakage. This indicates that the field of data science is interested and motivated by new work towards Leave-One-Out approaches that are not cross-validation and that this is a new area gaining traction. Although it is motivated by previous work such as ”K-fold cross validation” which certainly has far more work, the approach is distinct and contributes directly to the progress of the field of data science.

K-fold cross validation typically has the following properties that do not hold in this dissertation. First, k-fold cross validation has each fold approximately the same size [3]. In Chapter 5, the observation characteristics vary significantly. Some observations have 10 images while others can have up to 71 images. In fact, this wide variation is an extremely important aspect of deep phenotyping as it gives insight into why certain observations have more labelled data than others. One idea would be that the people having more labelled data have a more severe case with more cells needing to be detected. However, the opposite holds in this dissertation which makes k-fold cross validation a poor choice for this reason. In contrast, domain adaptation methods regularly deal with source and target domains of different sizes. In fact, one of the largest motivators for using a domain adaptation method is when the target domain is significantly smaller than the source domain[90]. Instead of needing to obtain more data in order to create an informative model, domain adaptation allows a model trained on a (typically much larger) source domain to be used on a given target domain. Second, k-fold cross validation struggles with class imbalances. He and Ma in 2013 state ”A 10-fold cross-validation, in particular, the most commonly used error-estimation method in machine learning, can easily break down in the case of class imbalances, even if the skew is less extreme than the one previously considered” in their book *Imbalanced Learning: Foundations, Algorithms, and Applications*[34]. In the case of medical image segmentation, large class imbalances are common since the cells of interest are typically the minority class in an image set. The underlying issue with k-fold cross validation on imbalanced classes is that the model will likely focus on accurately predicting the majority class. Given that the majority class in medical image segmentation consists of all the pixels that are not in the areas of interest, the k-fold cross validation would fail to yield any useful results for this research problem. In contrast, there have been domain adaptation methods have seen strong advances targeting problems with large class imbalances[37][84].

To evaluate different domain adaptation approaches, a full factorial statistical design is used. The three factors are domain size, domain choice, and threshold value. The evaluation metric will be the validation dice score. This dissertation will focus on a full factorial design targeting domain adaptation factors on the EoE dataset which are shown in Table 4.1.

Thus, the experiment will be an exploration of all possible combinations of these three factors. Continuing this example with the EoE Dataset would give the following design matrix shown in Table 4.2.

In this approach for the EoE dataset, the eight runs will provide a complete view

	Low (-1)	**High (+1)**
Domain Size	5	15
Domain Choice	Random	EoE/Not EoE
Threshold Value	0.3	0.5

Table 4.1: EoE Full Factorial Design: This experimental design will have two levels and three factors. The goal will be to explore the effect Domain Size, Domain Choice, and Threshold Value have on a multisource domain adversarial network dice score results.

Run	Domain Size	Learning Rate	Threshold Value
1	-	-	-
2	+	-	-
3	-	+	-
4	+	+	-
5	-	-	+
6	+	-	+
7	-	+	+
8	+	+	+

Table 4.2: EoE 2^3 Full Factorial Design Matrix: All factor combinations are tested to assess their effects on the multisource domain adversarial network dice score results.

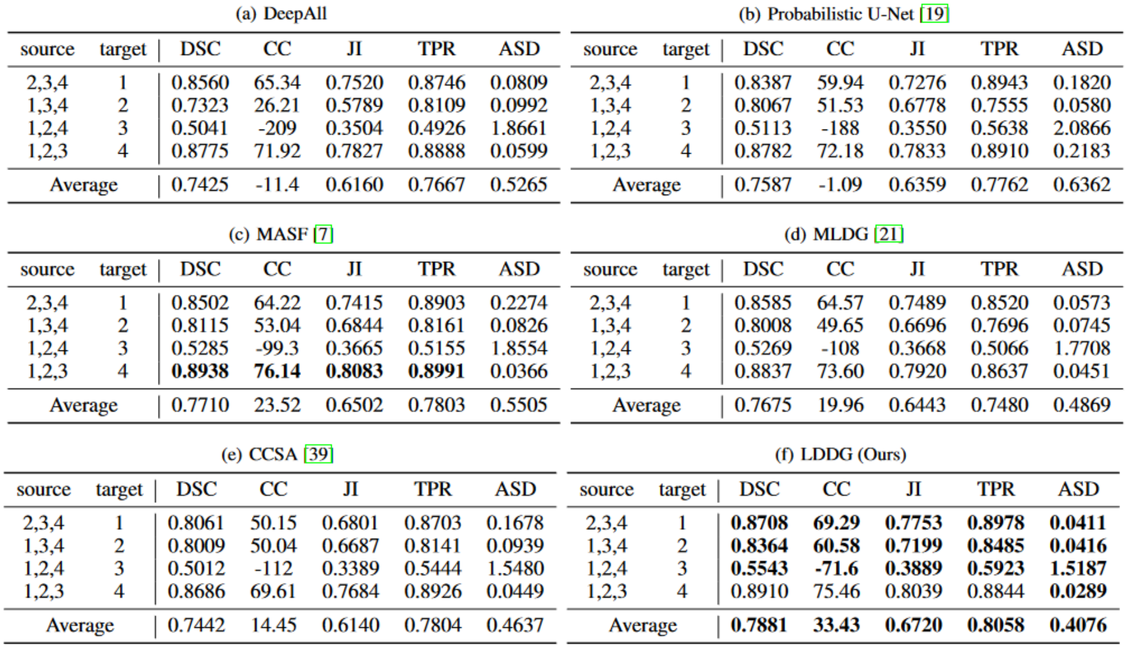
of the domain adaptation techniques results in response to changes in domain size, domain choice, and threshold value. Results would indicate whether these techniques can rival existing deep learning approaches and provide insight into how the model is learning to adapt to different domains.

4.2 Experiment Details

All experiments will be run using GPU nodes on Rivanna, UVA's high performance computer. For training, All models will be coded in Python preferably using the Keras API on TensorFlow back-end as a deep learning framework. If needed, PyTorch deep learning framework packages will also be used. For domain adaptation, there are several libraries that may assist with implementation to include: Awesome Domain Adaptation Python Toolbox (ADAPT) and Another Domain Adaptation Library (ADA). Additionally, there is a trade-of between segmentation performance and computation time. While the dice score for segmentation may continue to improve, eventually the overall optimization run time can become excessive. Excessive run times can hinder executing all necessary experiments in a timely fashion. The process that can produce an effective domain adaptation method within a reasonable time period should is preferred. An early stopping criteria was implemented to stop training when validation loss did not improve after ten consecutive epochs. The Adam optimizer was used to train model parameters.

4.3 Motivation for Domain Adaptation Approaches for Segmentation

Given that the size of an WSI can be 100,000x100,000px large [88], detecting diseases using medical images can be an extremely draining task. Recent medical imaging classification and segmentation efforts have attempted to leverage the strengths of deep learning to assist medical providers, identifying key areas for possible treatment and giving information about a patient's condition. Results have proven to be efficient and effective [32][75]. However, this progress has not properly addressed one of deep learning's most glaring issues: training. Traditional models require large amounts of data to train which may be difficult or impossible in the field of medical image diagnosis. Additionally, the training time needed for traditional models can grow significantly scaling dangerously with the size of the training dataset. Approaches such as domain adaptation have shown promising results by using source and target domains to minimize the training constraints. Continuing this progress, diffusion based methods have show impressive results, outperforming domain adaptation approaches in some studies [63]. Despite these advances in diffusion based methods, domain based methods have still shown comparable results, outperforming standard deep learning approaches and providing valuable information about specific domain choices [53]. For example, Li, et al from the Nanyang Technological University in Singapore used a domain adaptation approach to target better generalization capability to the "unseen" medical data. Li et al's approach used two different types datasets: skin lesion and spinal cord gray matter and evaluated domain adaptation methods using the following performance metrics. Dice Similarity Coefficient (DSC), Jaccard Index (JI), Conformity Coefficient (CC), True Positive Rate (TPR), and Average



(a) DeepAll

source	target	DSC	CC	JI	TPR	ASD
2,3,4	1	0.8560	65.34	0.7520	0.8746	0.0809
1,3,4	2	0.7323	26.21	0.5789	0.8109	0.0992
1,2,4	3	0.5041	-209	0.3504	0.4926	1.8661
1,2,3	4	0.8775	71.92	0.7827	0.8888	0.0599
Average		0.7425	-11.4	0.6160	0.7667	0.5265

(b) Probabilistic U-Net [19]

source	target	DSC	CC	JI	TPR	ASD
2,3,4	1	0.8387	59.94	0.7276	0.8943	0.1820
1,3,4	2	0.8067	51.53	0.6778	0.7555	0.0580
1,2,4	3	0.5113	-188	0.3550	0.5638	2.0866
1,2,3	4	0.8782	72.18	0.7833	0.8910	0.2183
Average		0.7587	-1.09	0.6359	0.7762	0.6362

(c) MASF [7]

source	target	DSC	CC	JI	TPR	ASD
2,3,4	1	0.8502	64.22	0.7415	0.8903	0.2274
1,3,4	2	0.8115	53.04	0.6844	0.8161	0.0826
1,2,4	3	0.5285	-99.3	0.3665	0.5155	1.8554
1,2,3	4	**0.8938**	**76.14**	**0.8083**	**0.8991**	0.0366
Average		0.7710	23.52	0.6502	0.7803	0.5505

(d) MLDG [21]

source	target	DSC	CC	JI	TPR	ASD
2,3,4	1	0.8585	64.57	0.7489	0.8520	0.0573
1,3,4	2	0.8008	49.65	0.6696	0.7696	0.0745
1,2,4	3	0.5269	-108	0.3668	0.5066	1.7708
1,2,3	4	0.8837	73.60	0.7920	0.8637	0.0451
Average		0.7675	19.96	0.6443	0.7480	0.4869

(e) CCSA [39]

source	target	DSC	CC	JI	TPR	ASD
2,3,4	1	0.8061	50.15	0.6801	0.8703	0.1678
1,3,4	2	0.8009	50.04	0.6687	0.8141	0.0939
1,2,4	3	0.5012	-112	0.3389	0.5444	1.5480
1,2,3	4	0.8686	69.61	0.7684	0.8926	0.0449
Average		0.7442	14.45	0.6140	0.7804	0.4637

(f) LDDG (Ours)

source	target	DSC	CC	JI	TPR	ASD
2,3,4	1	**0.8708**	**69.29**	**0.7753**	**0.8978**	**0.0411**
1,3,4	2	**0.8364**	**60.58**	**0.7199**	**0.8485**	**0.0416**
1,2,4	3	**0.5543**	**-71.6**	**0.3889**	**0.5923**	**1.5187**
1,2,3	4	0.8910	75.46	0.8039	0.8844	**0.0289**
Average		**0.7881**	**33.43**	**0.6720**	**0.8058**	**0.4076**

Figure 4.4: Li, et al.'s [53] Adaptation Method Performance Evaluation: Comparison of Domain Adaptation Approaches with other models using Dice Score (DSC), Jaccard Index (JI), Conformity Coefficient (CC), True Positive Rate (TPR), and Average Surface Distance (ASD) as performance metrics.

Surface Distance (ASD). The results are shown in Figure 4.4.

As with many other medical datasets, the concern for the UVA Medical Center EoE data is that it may not be sufficient for training and any model trained on the datasets can have issues with model generalization. The dataset only contains 30 patients and 514 labeled images/masks, motivating the effort to not use traditional deep learning approaches. Before importing the data into a model and extracting the features, the image distributions of all 30 patients are compared and visualized through a network diagram shown in Figure 4.5. The network is shown with the most unique patients towards the edges of the diagram and the most unique patients in the center. Although all patients were ultimately diagnosed withe EoE at UVA Medical Center, the network color codes the patients who's labeled images/masks would meet the "15 or more eosinophil" diagnostic criteria. Additionally, the entropy for each patient is represented this through the circle sizes. This was done through a Monte Carlo Dropout UNet approximation to the deep Gaussian process used in Bayesian Neural Networks [27]. Minimizing the KL divergence using an approximation through Monte Carlo integration to gives an unbiased estimator. The entropy [28] is then given by the following equation.

$$E_{p(z|D)}H[p(y|z,x)] = -\int p(z|D)\left(\sum_{y\in Y} p(y|z,x)\log p(y|z,x)dw\right) \tag{4.2}$$

Immediately apparent was that there were some serious differences in patient samples which are hereby defined as observations. Two quick examples were that observation E-139 had more than 50 times the entropy of observation E-28 and that observation E-17 had 71 labelled biopsy images while most observations had only 10 labelled biopsy images. Overall, it also can be concluded that the higher the entropy, the less unique an observation is relative to the group. These significant differences gives an indication that the dataset is far from homogenous and that traditional deep learning techniques may struggle to detect areas of interest. Thus, domain adaptation and diffusion approaches are a natural progression from the information gathered in the deep dive into the observations and from advancements compared to standard deep learning techniques. This analysis presents progression in both domain adaptation and diffusion based approaches in medical imaging.

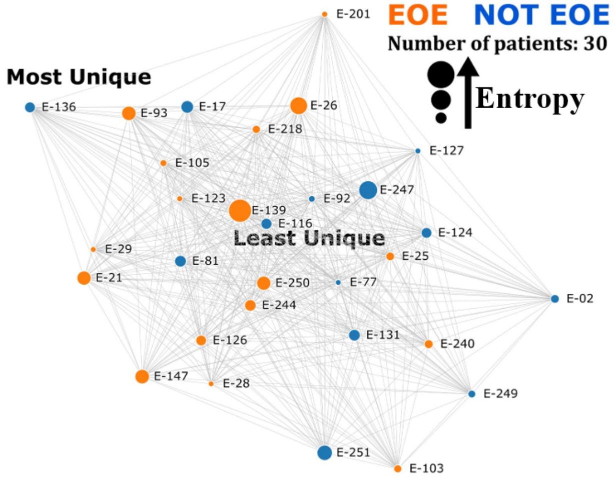


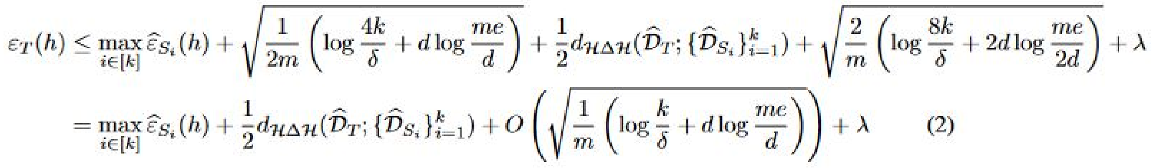
Figure 4.5: EoE Dataset Network Diagram: Observation-based visualization shows that there is significant variation between each patient's dataset that standard deep learning approaches struggle to address.

4.4 Domain Adaptation and Diffusion Based Methods[57]

For the domain adaptation approach, the dataset is split into three different domains: low, medium, and high entropy observations. Since there are 30 observations, each domain will have 10 observations. Each of the combinations is then trained and evaluated for average performance calculated through a Fréchet Inception Distance (FID), precision, and recall. Letting D_T be the target domain and D_{S_i} be the source domain over X, then the generalization bound is given by

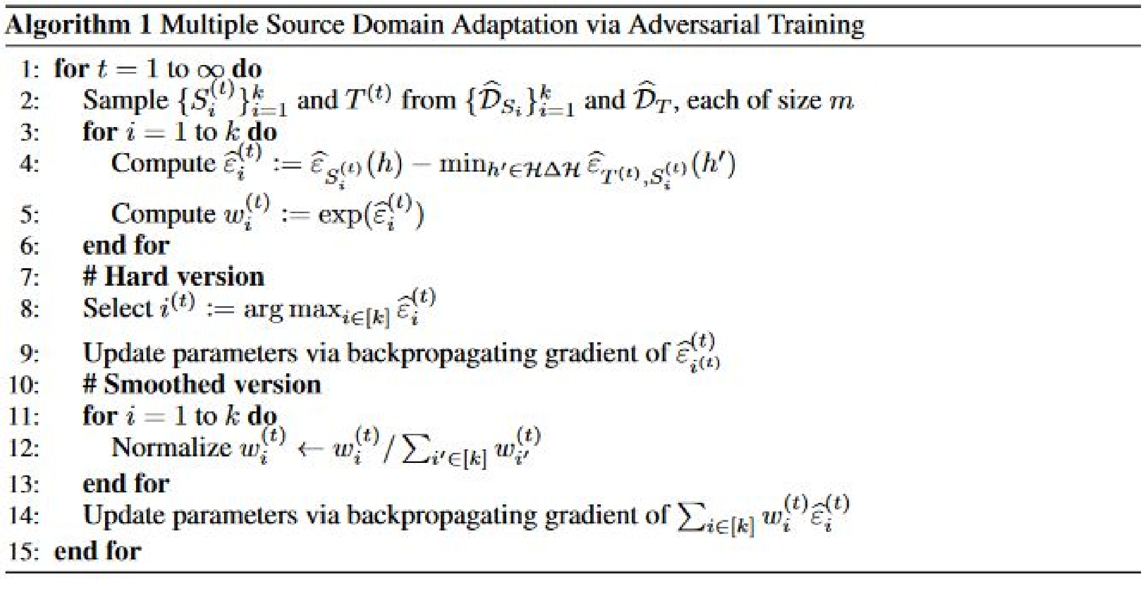
The goal in the multisource domain adversarial network approach is to minimize the generalization bound shown in Figure 4.6. This is typically done through a minimax saddle point problem and then optimized through adversarial learning

Diffusion based models use variational inference to produce samples matching the



$$\begin{aligned}\varepsilon_T(h) &\le \max_{i\in[k]} \widehat{\varepsilon}_{S_i}(h) + \sqrt{\frac{1}{2m}\left(\log\frac{4k}{\delta} + d\log\frac{me}{d}\right)} + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\widehat{\mathcal{D}}_T; \{\widehat{\mathcal{D}}_{S_i}\}_{i=1}^{k}) + \sqrt{\frac{2}{m}\left(\log\frac{8k}{\delta} + 2d\log\frac{me}{2d}\right)} + \lambda \\ &= \max_{i\in[k]} \widehat{\varepsilon}_{S_i}(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\widehat{\mathcal{D}}_T; \{\widehat{\mathcal{D}}_{S_i}\}_{i=1}^{k}) + O\left(\sqrt{\frac{1}{m}\left(\log\frac{k}{\delta} + d\log\frac{me}{d}\right)}\right) + \lambda \qquad (2)\end{aligned}$$

Figure 4.6: Multisource Domain Adversarial Network Generalization Bound: Combines the worst classification accuracy, distance between the source and target domains, optimal error, and data bias into a generalization bound



Algorithm 1 Multiple Source Domain Adaptation via Adversarial Training

```
for $t = 1$ to $\infty$ do
    Sample $\{S_i^{(t)}\}_{i=1}^{k}$ and $T^{(t)}$ from $\{\widehat{\mathcal{D}}_{S_i}\}_{i=1}^{k}$ and $\widehat{\mathcal{D}}_T$, each of size $m$
    for $i = 1$ to $k$ do
        Compute $\widehat{\varepsilon}_i^{(t)} := \widehat{\varepsilon}_{S_i^{(t)}}(h) - \min_{h' \in \mathcal{H}\Delta\mathcal{H}} \widehat{\varepsilon}_{T^{(t)}, S_i^{(t)}}(h')$
        Compute $w_i^{(t)} := \exp(\widehat{\varepsilon}_i^{(t)})$
    end for
    # Hard version
    Select $i^{(t)} := \arg\max_{i\in[k]} \widehat{\varepsilon}_i^{(t)}$
    Update parameters via backpropagating gradient of $\widehat{\varepsilon}_{i^{(t)}}^{(t)}$
    # Smoothed version
    for $i = 1$ to $k$ do
        Normalize $w_i^{(t)} \leftarrow w_i^{(t)} / \sum_{i'\in[k]} w_{i'}^{(t)}$
    end for
    Update parameters via backpropagating gradient of $\sum_{i\in[k]} w_i^{(t)} \widehat{\varepsilon}_i^{(t)}$
end for
```

Figure 4.7: Multisource Domain Adversarial Network Algorithm: Stores informative feature representations and captures invariant feature representations between different domains.[8]

data given sufficient time. This is typically done through a parameterized Markov chain that gradually adds noise to the image until the signal is destroyed. A common checkpoint in these approaches is to make sure that the noise indeed reduces the signal to noise ratio to zero (or close to it). In comparison to other deep learning techniques, diffusion models are straightforward to define and efficient to train but there has been limited demonstration that they are capable of generating high quality samples. The following implementation shown in Algorithm 1 will be used to apply this diffusion based approach to the EoE dataset[36]. This baseline approach will also be used to compare FID, precision, and recall for both the multisource domain adversarial network and diffusion approaches. A Monte Carlo Dropout UNet with a dropout value set to 0.5 will be the baseline model for this work. All training was done on 4 NVIDIA A100 GPUs with 300GB of RAM in TensorFlow/Keras 2.7/PyTorch 2.1.1. Each model run was ran for 100 epochs with a learning rate of 0.001. For the domain adaptation and the diffusion based models, the evaluation metric will be a Fréchet inception distance comparing the generated and real images. For all three models, precision and recall will also be recorded.

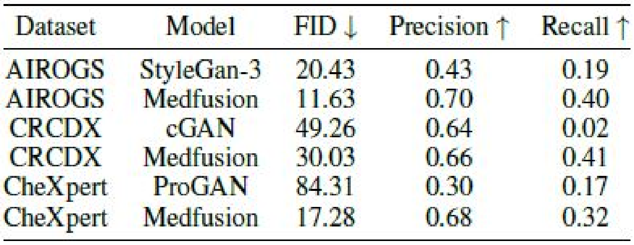
Algorithm 1 Training

```
repeat
$x_0 \sim q(x_0)$
$t \sim \text{Uniform}(1, \ldots, T)$
$\epsilon \sim N(\mathbf{0,I})$
Take gradient descent step on
    $\nabla_\theta||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}}x_0 + \sqrt{1-\bar{\alpha}}\epsilon, t)||^2$
until converged
```

Although there are a few papers [63] stating that the Diffusion Probabilistic Models beat GANs on medical 2-D images, the results in Table 4.3 suggest that there is at least a space where the GANs may capture feature representations better than diffusion methods. In addition, these papers also have cases where some GANs had

Model	FID	Precision	Recall
MCD UNet	N/A	0.567	0.495
MDAN	42.56	0.623	0.657
DDPM [36]	50.65	0.489	0.457

Table 4.3: Model Performance Comparison: Multisource domain adversarial network approach (MDAN) outperforms diffusion methods (DDPM) on EoE Dataset.



Dataset	Model	FID ↓	Precision ↑	Recall ↑
AIROGS	StyleGan-3	20.43	0.43	0.19
AIROGS	Medfusion	11.63	0.70	0.40
CRCDX	cGAN	49.26	0.64	0.02
CRCDX	Medfusion	30.03	0.66	0.41
CheXpert	ProGAN	84.31	0.30	0.17
CheXpert	Medfusion	17.28	0.68	0.32

Figure 4.8: Diffusion vs GAN [63]: Diffusion methods (Medfusion) have shown strong performance against adversarial methods (StyleGan-3, cGAN, ProGAN) but some results are comparable such as the precision of cGAN vs the precision of Medfusion.

performance comparable to diffusion methods as shown in Figure 4.8. In this table, cGAN has a precision of 0.64 which is close to Medfusion's precision of 0.66. With a GAN approach achieving similar results as a diffusion approach, the results observed from the multisource domain adversarial network approach in Table 4.3 motivate future works exploring the trade-space between adversarial and diffusion approaches. Furthermore, the precision and recall metrics for the domain adaptation approach are better than both the diffusion and the baseline UNet approach.

As a check for the diffusion based methods actually eliminating the signal from a given image, the EoE images after diffusion shown are plotted in Figure 4.9. Clearly, the signal is completely destroyed and there is no indication of cellular structures anywhere in the images. Therefore, the fact that the diffusion based method was able to extract information from this image being in such a state is impressive.

The strong performance of the domain adaptation methods compared to the diffusion

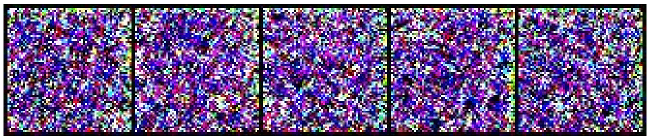


Figure 4.9: Destruction of the EoE Signal: Denoising Diffusion methods (DDPM) require a destruction of an image's signal by adding noise as an intermediary step before generating synthetic images.

and the standard UNet approach motivates further work in the domain adaptation field. Next steps could be to see which of the three domains used were most impactful to the relatively strong performance and what impact, if any, would their removal from the dataset be. Additionally, one of the advantages in using a domain adaptation approach is exploring the relationships between different source and target domains. Adding different datasets such as the public PanNuke or Lizard dataset to the analysis can provide different insights into the EoE patients here at UVA and potentially help other researchers in the field of medical image diagnosis.

4.5 Extremity-Ranked Domain Selection (ERDS)

Deep learning has been well established as an appropriate and effective approach for medical image segmentation[91]. A major limitation of these deep learning approaches is the lack of generalization and explainability. Recent approaches to address these issues are the addition of domain adaptation methods that serve to give insight into how different models trained on other datasets can assist future efforts to create new models[105]. The benefits of effectively applied domain adaptation techniques span multiple research fields including image captioning[17], object detection[87], and medical image segmentation[79]. Out of the these fields, medical image segmentation has shown rapid progress with UNet[2], diffusion[63], and GAN[95] approaches showing strong abilities in detecting cells of importance. However, the core focus of medical work is observation-centric with improving patient care a number one priority for medical providers and these above methods fail to address observation-level assessment in their results. In contrast, Extremity Ranked Domain Selection (ERDS) focuses on the impact each observation has on the model's performance. ERDS transverses the entire dataset, individually removing an observation from the training data, calculates the extremity by assessing how much a baseline MCD UNet dice score decreases when they are removed, and then ranks the observation by their extremity. This task yields valuable information on how deep learning models rely on subsets of training data and how certain subsets of training data may be sufficient enough to span all necessary features. With these observation rankings, domain selection begins with creating multiple domains of high and low observation extremity. To assess these domains, a full factorial experiment is created spanning domain size, domain choice (random or extremity ranked), and classification threshold value. Ultimately, ERDS yields the optimal values in the full factorial experiment, demonstrates that the value of this extremity metric in future works, and motivates further research in analyzing impacts of varying training data.

Chapter 5

Results

5.1 Baseline MCD Model

Prior to incorporating a new deep learning approach, baseline models must be created to not only produce comparable dice scores but also as a check to ensure that results are reasonable. In this work, the Monte Carlo Dropout UNet [71] will be the baseline model. All training in this work will be using 100 epochs, ADAM optimizer, TensorFlow 2.13.0, 384 GB of storage, and 4 NVIDIA A100 GPUs. Evaluation of the work will be through the entropy metric calculated in Equation 4.2 and a Dice Score. For the baseline model, the results of using this approach are shown in Figure 5.1. To avoid issues with the model outputting high dice scores when training on "blank" patches, any images with a Dice Score of exactly 1.0 or 0.0 are removed. Performance of a Monte Carlo UNet yields a median of 0.576 for the dice score which matches current medical image segmentation research results[2]. To ensure that this baseline model is performing sufficiently well, a ROC Curve is calculated and shown in Figure 5.2. The ROC Curve verifies that the model has is appropriate to use as a baseline with some values closely approaching the top left corner of the graph and an AUC of 0.971. This indicates that the classification accuracy of the model is high and clearly better than a true random guess. However, the EoE dataset has severe class imbalances with the majority of pixels in an image patch not in the eosinophil class.

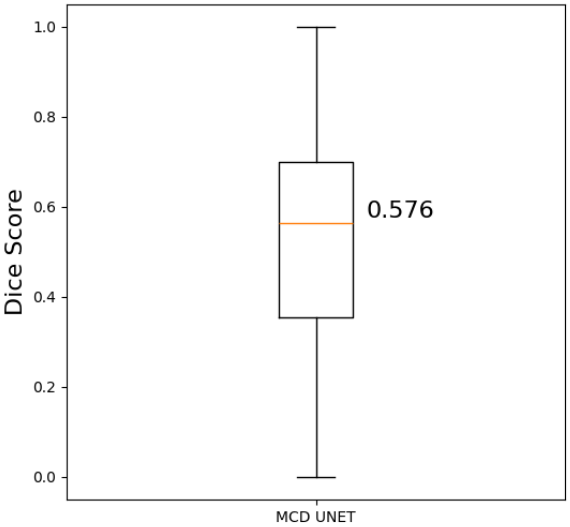


Figure 5.1: Baseline model dice scores: Monte Carlo Dropout UNet. The results are reasonable given previous results shown in Figure 3.6.

In fact, EoE pixels numbered 2,339,280 which is approximately 2% of all pixels while non-EoE pixels numbered 111,431,116 which is approximately 98% of all pixels. For this reason, the Dice Score is a far more appropriate evaluation metric. With these strong preliminary results, ERDS can proceed by using MCD UNet as an appropriate baseline model.

5.2 ERDS Approach

Extremity-Ranked Domain Selection represents a combination of new domain adaptation techniques with statistical approaches to determine the effect of modifying a training dataset on a MCD UNet dice score. Figure 5.3 illustrates a schematic of the approach. First, a baseline run is conducted with a Monte Carlo UNet and the full training dataset. Then, observations are removed with replacement from the training

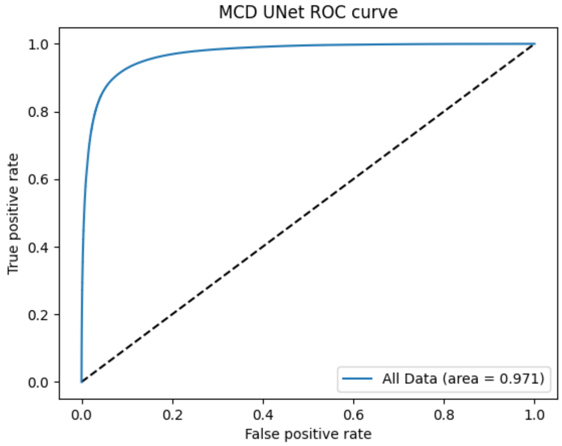


Figure 5.2: MCD ROC Curve: EoE pixels = 2,339,280 (≈2%), non-EoE pixels = 111,431,116 ($approx$98%). Large class differences motivate using Dice Score as an evaluation metric.

dataset and the same Monte Carlo UNet is run. With each run, the dice score and entropy are recorded for each observation. Observations are then ranked on their extremity which is determined by their median dice score and median entropy values. Domains are created by grouping observations of high extremity and low extremity. This creates the source domains needed for the Multisource Domain Adversarial Networks (MDAN)s. Although the domains use observations from the same EoE dataset, the observation characteristics are not the same and create the difference needed in order to define different source and target domains. To be clear, this research is not focused on evaluating on multiple diseases or multiple datasets. This research is concerned about how EoE presents over this set of observations from the UVA Medical Center. Given that each observation has unique characteristics including age, sex, ethnicity, BMI, location of biopsy taken, and time of initial biopsy collection, this approach will provide information for deep phenotyping efforts in UVA's eosinophil detection efforts. In addition to different medical characteristics, observations also have wide dataset variations. For example, observation E-139 had zero eosinophils in 28 out of 41 of their patches while observation E-77 had so many eosinophils their patches that the number of pixels that had eosinophils approached the number of pixels that did not have eosinophils. E-77 specifically seemed to have nearly balanced class sizes in their input patches.

Finally, a full factorial experimental design is created focusing on domain size (5, 15), domain choice (Random, Extremity Ranked), and threshold value (0.3, 0.7). For the target domain, one observation is randomly chosen in these full factorial runs to be the target domain. The benefit of using a full factorial experiment is that each of these factors domain size, domain choice, and classification threshold are assessed to determine their respective effects on the MDAN dice score. Compared to other approaches, the full factorial design also accounts for interaction terms combining these factors to see if a subset of them drive the model's dice score. These results can be visualized through a standard least squares approach focusing on effect screening yielding information as to if these factors are an appropriate choice for predicting MDAN dice score. Finally, optimal combinations of these factors can be calculated, demonstrating which factors actually drive model dice score and driving future research into sensitivity analysis for these factors. All full factorial design experiments are conducted on JMP using the Student Licenses.

Before starting, the EoE dataset has some slight class imbalances shown in Table 5.1. In particular, observations E-17, E-105, and E-139 have significantly more patches of 71, 63, and 41 patches respectively in comparison to most other observations who have only 10 patches. Discussion with the medical researchers at the University of Virginia Gastroenterology Data Science Laboratory showed that some of these patients visit the hospital multiple times for these biopsy collections and that some patients may have multiple areas of interest in their WSI. Given that each of these

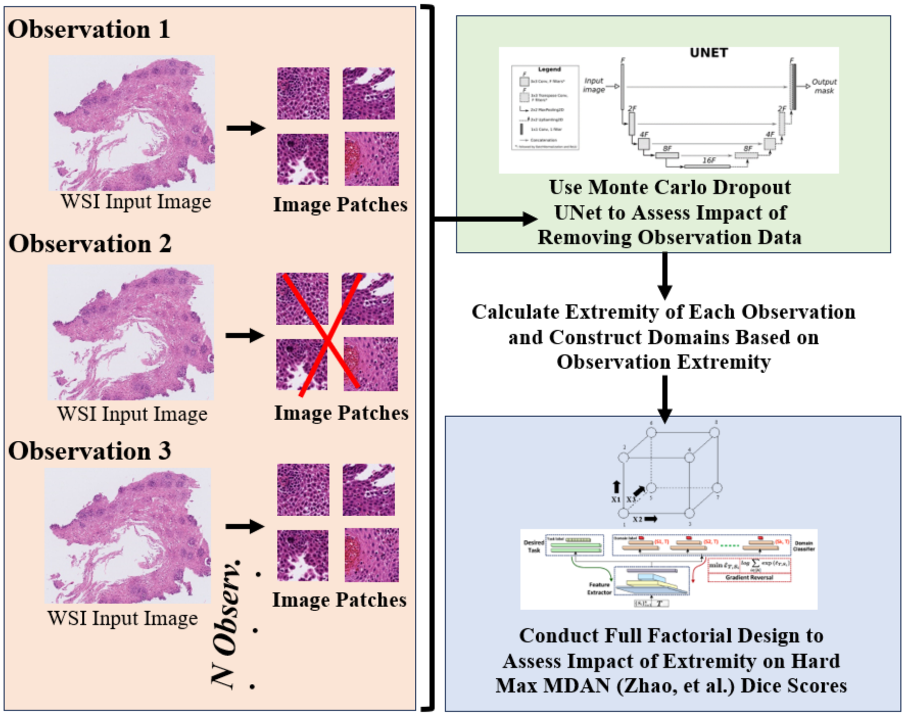

Observation 1
WSI Input Image
Image Patches
Observation 2
WSI Input Image
Image Patches
Observation 3
WSI Input Image
Image Patches
N Observ.
UNET
Use Monte Carlo Dropout UNet to Assess Impact of Removing Observation Data
Calculate Extremity of Each Observation and Construct Domains Based on Observation Extremity
X1
X3
X2
Conduct Full Factorial Design to Assess Impact of Extremity on Hard Max MDAN (Zhao, et al.) Dice Scores

Figure 5.3: Extremity Ranked Domain Selection Approach. Model dice score is tested by individually removing observations from the training dataset prior to training the model, creating a extremity metric. Observations are then ranked by their extremity and placed into a domain. A Full Factorial Design of Experiment (DOE) using dice scores from a multisource domain adversarial networks (MDAN) using the observations as appropriate source and target domains.

patches are traditionally manually labeled by medical pathologist, observation E-17 with 71 patches would be more resource intensive for disease detection than E-02 with only 10 patches. This seven-fold different in observation class sizes can lead to different labeling results and motivates the ERDS research of assessing the model dice score after iteratively removing each observation. Lastly, this indicates that the observations with higher amounts of patches may have a larger eosinophil distribution in their WSIs which can lead to more difficulties for medical pathologists to diagnose. Given that the size of a WSI can be up to 100,000 x 100,000 px and that the patches are 512x512 px, selecting appropriate patches to label can be significantly resource heavy. In order to correct these class imbalances, observations will be weighed by their amount of images in the dataset.

The extremity results from running the Monte Carlo Dropout UNet and removing each observation individually are shown in Figures 5.4 and 5.5. The full dataset results are indicated by a thick horizontal line for comparative analysis. Extremity for this MDAN approach will be defined as the median dice score instead of the mean dice score to prevent significant impact from any extreme observations. Formally, let X be the set of all observations in the dataset, $D(X_i)$ represent the dice score of a MCD UNet trained on all data except for observation i, and $Mdn()$ represent the median. The patient observation extremity (POE) metric is then given by:

$$\text{Patient Observation Extremity}(i) = |Mdn(D(X)) - Mdn(D(X_i))| \tag{5.1}$$

Verbally, extremity is then the non-negative difference between a baseline Monte Carlo Dropout UNet dice score with all observations in the training dataset and a baseline Monte Carlo Dropout UNet dice score with a specific observation removed.

The graphs also demonstrate that there is not a significant difference between choosing median or mean dice score as a factor in the extremity ranking metric. As expected, removing a observation lowers the already relatively small dataset which almost universally negatively impacts the dice score metric. In fact, the MCD UNet dice score only improved when observations were removed for ten patients. Removing observation E-139 had the largest increase in dice score of 0.65, outperforming the baseline model dice score of 0.62. Correspondingly, removing observation E-139 also had the smallest entropy of 0.0027 compared to the baseline MCD UNet entropy of 0.014 demonstrating that this observation seems to hinder model dice score. Referring back to the dataset distribution table shown in Table 5.1, observation E-139 has 41

Observation Name	Number of Patches	Percentage of Dataset
E-02	10	1.95%
E-17	71	13.81%
E-21	24	4.67%
E-25	33	6.42%
E-26	15	2.92%
E-28	10	1.95%
E-29	10	1.95%
E-77	10	1.95%
E-81	10	1.95%
E-92	10	1.95%
E-93	10	1.95%
E-103	10	1.95%
E-105	63	12.26%
E-116	37	7.20%
E-123	10	1.95%
E-124	10	1.95%
E-126	10	1.95%
E-127	10	1.95%
E-131	10	1.95%
E-136	10	1.95%
E-139	41	7.98%
E-147	10	1.95%
E-201	10	1.95%
E-218	10	1.95%
E-240	10	1.95%
E-244	10	1.95%
E-247	10	1.95%
E-249	10	1.95%
E-250	10	1.95%
E-251	10	1.95%

Table 5.1: EoE Patient Data Distribution: Wide variation in number of patches per observation indicate that a more observation-centric approach to deep learning is appropriate for this EoE dataset. Standard deep learning approaches would not account for observation-level analysis.

patches which is the third largest patient dataset. Visual analysis of the patches for observation E-139 showed zero eosinophils for a majority of the patches and is shown in Table 5.2. The patches in observation E-139 lack eosinophils which means that these patches will provide limited information for training the model to detect and identify eosinophils. The lack of eosinophils in observation E-139's patches also explains why the entropy actually decreased when observation E-139 was removed from the training dataset. The addition of blank true patches only caused the model to become more stable because the model is already trained on data with large amount of class imbalances. Remember that the majority of the pixels present in input images do not correspond to the eosinophils and are considered part of the background class. To connect this class balance visually, most pixels in the input image are black (background) instead of white (eosinophils). Nearly all pixels in observation E-139's patches were background pixels. Therefore, adding more background pixels to a training set already heavily imbalanced towards background pixels will reduce the entropy. Discussion with the UVA Gastroenterology Data Science Lab researchers yielded the following characteristics for observation E-139. Sex: M, Ethnicity: Non-Hispanic, Race: White, Age 13, BMI: 28.58. Observation E-139 was also already diagnosed with EoE prior to biopsy. Observation E-139 likely received treatment for EoE which would explain the relative lack of eosinophils present in the patches. Because some patients were already diagnosed with EoE before their biopsy, the deep phenotyping efforts may be difficult due to the treatment's effect on the observations.

Conversely, removing observation E-123 had the largest decrease in dice score of 0.287 and correspondingly had the highest entropy of 0.057 demonstrating that this observation seems to be crucial to obtaining a strong MCD UNet dice score on medical image segmentation on the EoE dataset. Table 5.1 only shows 10 patches for observa-

Number of Eosinophils	Number of Patches
0	28
1	7
2	4
3	1

Table 5.2: Number of eosinophils present in each patch for observation E-139. The majority of patches (28/41) have zero eosinophils indicating that the data from observation E-139 has a severe class imbalance which will impact the model's ability to learn how to detect and identify eosinophils.

tion E-123 indicating that the features in this observation's data seem to completely span the necessary features needed for the MCD UNet to detect eosinophils. Visual analysis of the patches for observation E-123 yielded wide variation in eosinophil location and orientation which would be necessary to inform a deep segmentation model trained to detect and identify eosinophils. Discussion with the UVA Gastroenterology Data Science Lab researchers yielded the following characteristics for observation E-123. Sex: F, Ethnicity: Non-Hispanic, Race: White, Age 22, BMI: 22.6. Observation E-123 also was diagnosed with EoE prior to biopsy. Interestingly, observation E-123 still has significant amounts of eosinophils in their patch which means that treatment may not be as effective in reducing the amount of eosinophils present. One observation of interest is E-93 who had a relatively average dice score but extremely high entropy when removed indicating that their data has some stabilizing effect on the MCD UNet dice score. This stabilizing effect does not seem to translate to the MCD UNet dice score indicating that some of the dice score results may vary significantly between runs and different datasets. However, this variation is minimized through taking multiple runs and taking the median dice scores over the runs. Discussion with the UVA Gastroenterology Data Science Lab researchers yielded the following characteristics for observation E-93. Sex: F, Ethnicity: Non-Hispanic, Race: White, Age 46, BMI: 33.23. In contrast with the patients discussed before, observation E-93 did not have EoE prior to biopsy. This means that the disease progression most likely has not been mitigated by treatment and this observation has a higher BMI than most other observations. These factors can explain part of why there is high amounts of entropy when this observation is removed from the training dataset. A final observation of interest is observation E-77 who had the lowest POE value. Visual inspection of E-77's patches showed that they differed considerably from other observations. Whereas most observations have a serious class imbalance with the minority of pixels associated with eosinophils, the patches of obseration E-77 were actually rife with eosinophils with an average of over 30 eosinophils per patch. This means that for this observation, the amount of pixels that were associated with the background and the amount of pixels associated with the eosinophils were roughly equivalent. Therefore, removing observation E-77 from the model had little impact because this observation had data with no class imbalances which differs from the other observations' patches. Discussion with the UVA Gastroenterology Data Science Lab researchers yielded the following characteristics for observation E-77. Sex: M, Ethnicity: Non-Hispanic, Race: White, Age 35, BMI: 22.96. Observation E-77 was also not diagnosed with EoE prior to biopsy. Given the significantly larger amounts of eosinophils present in observation E-77's patches, this observation's EoE has also most likely not been mitigated by treatment. However, observations E-77's differences in characteristics make removing this observation in training have a less effect than observation E-93.

With these results, the POE metric is now defined as the distance between a removed observation's median MCD UNet dice score with the baseline MCD UNet dice score and observation extremity is shown in Table 5.3 and visualized in Figure 5.6. This

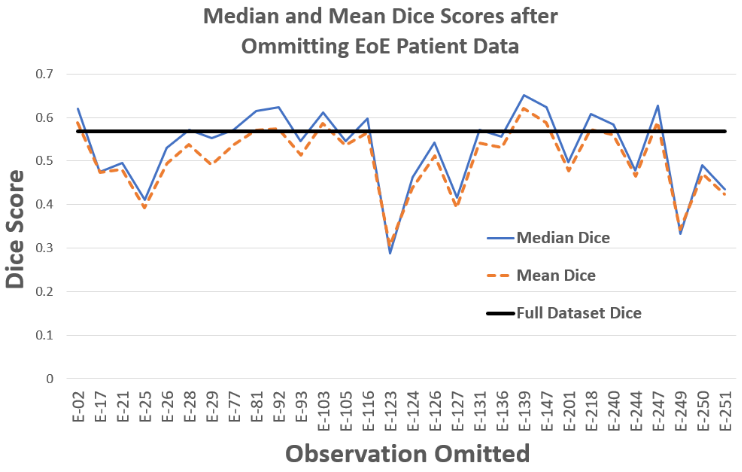


Figure 5.4: Median and Mean Dice Scores: Removing an observation's training data negatively impacts the dice score in all but ten cases. Removing observation E-139 had the largest increase in dice score while removing observation E-123 had the largest decrease in dice score.

distance is non-negative to account for both the positive and negative effects shown by removing an observation's data. The top five observationss for extremity: E-123, E-249, E-25, E-127, and E-251 all negatively impacted the MCD UNet when their training data was removed. The observation with the highest extremity that positively impacted the MCD UNet was observation E-139 who was ranked ninth highest in extremity. This indicates that in medical image segmentation, the lack of a certain observation's training data is more likely to have a significantly negatively impact on model dice score than a positive impact.

These extremity results yielded the following domain selections for two, shown in Table 5.4 and six domains, shown in Table 5.5. For the two domain approach in Table 5.4, Domain 1 represents observations with high extremity while Domain 2 represents observations with low extremity. For the six domain approach in Table 5.5, Domain 1-3 represents observations with high extremity while Domain 4-6 represents observations with low extremity. With the source domains created, the full factorial

Observation Name	Number of Patches	Percentage of Dataset	POE
E-02	10	1.95%	0.051051
E-17	71	13.81%	0.093689
E-21	24	4.67%	0.074188
E-25	33	6.42%	0.157987
E-26	15	2.92%	0.038915
E-28	10	1.95%	0.002608
E-29	10	1.95%	0.016137
E-77	10	1.95%	0.002523
E-81	10	1.95%	0.045472
E-92	10	1.95%	0.053975
E-93	10	1.95%	0.022555
E-103	10	1.95%	0.041581
E-105	63	12.26%	0.023355
E-116	37	7.20%	0.029147
E-123	10	1.95%	0.281666
E-124	10	1.95%	0.106203
E-126	10	1.95%	0.026039
E-127	10	1.95%	0.152962
E-131	10	1.95%	0.002639
E-136	10	1.95%	0.013669
E-139	41	7.98%	0.081651
E-147	10	1.95%	0.055281
E-201	10	1.95%	0.071265
E-218	10	1.95%	0.039123
E-240	10	1.95%	0.015329
E-244	10	1.95%	0.08995
E-247	10	1.95%	0.058182
E-249	10	1.95%	0.235778
E-250	10	1.95%	0.079471
E-251	10	1.95%	0.133754

Table 5.3: EoE Observation Extremity: The top eight observations with the highest extremity had a negative impact on the dice score when their EoE data was removed in training. Observation E-123 had the highest extremity while observation E-77 had the lowest extremity.

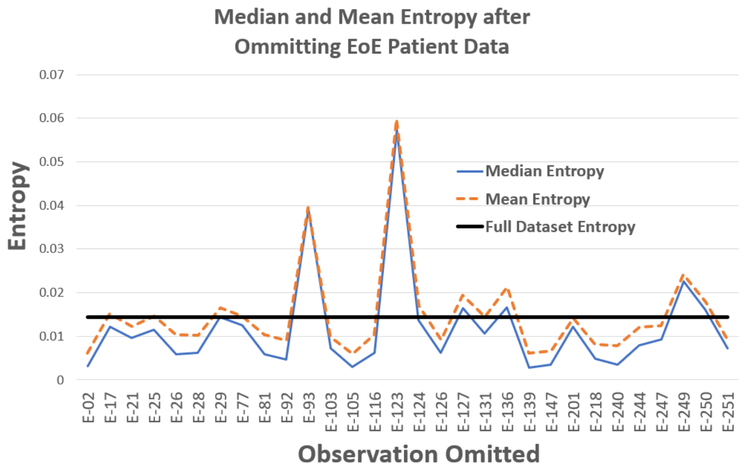


Figure 5.5: Median and Mean Entropy: Removing an observation's training data decreases the entropy in all but nine cases. Removing observation E-23 had the largest increase in entropy while removing observation E-139 had the largest decrease in entropy.

design and MDAN approach can be executed.

The full factorial design represents a robust statistical method to determine optimal parameter estimates. The full factorial design used consists of three factors and two levels, yielding eight total runs. The factors and levels are domain size (5, 15), domain choice (Random, Extremity Ranked), and classification threshold (0.3, 0.7). Using the MDAN, the full factorial design provides the following results shown in Figures 5.7 and Figure 5.8. The strong R^2 value of 0.9845 and strong linear correlation on the standard least squares output indicate that the choice of factors for this model is appropriate. Furthermore, the graphs at the top of Figure 5.8 are 2-d representations equivalent to a response surface and illustrate that the optimal combination of factors to maximize the MDAN dice score are domain size of 15, domain choice of Extremity Ranked, and classification threshold of 0.3. Finally, the JMP output displays a desirability metric that maximizes the geometric average of multiple two-sided transformations of each response [4]. The desirability of the optimal factor combination is shown to be

Domain 1	Domain 2
E-17	E-02
E-21	E-26
E-25	E-28
E-92	E-29
E-123	E-77
E-124	E-81
E-127	E-93
E-139	E-103
E-147	E-105
E-201	E-116
E-244	E-126
E-247	E-131
E-249	E-136
E-250	E-218
E-251	E-240

Table 5.4: Two Domain Division: Domain 1 consists of observations with high extremity. Domain 2 consists of observations with low extremity.

Domain 1	Domain 2	Domain 3	Domain 4	Domain 5	Domain 6
E-25	E-17	E-21	E-02	E-29	E-28
E-123	E-124	E-92	E-26	E-93	E-77
E-127	E-139	E-147	E-81	E-105	E-131
E-249	E-244	E-201	E-103	E-116	E-136
E-251	E-250	E-247	E-218	E-126	E-240

Table 5.5: Six Domain Division: Domains 1-3 consist of observations with high extremity. Domains 4-6 consist of observations with low extremity.

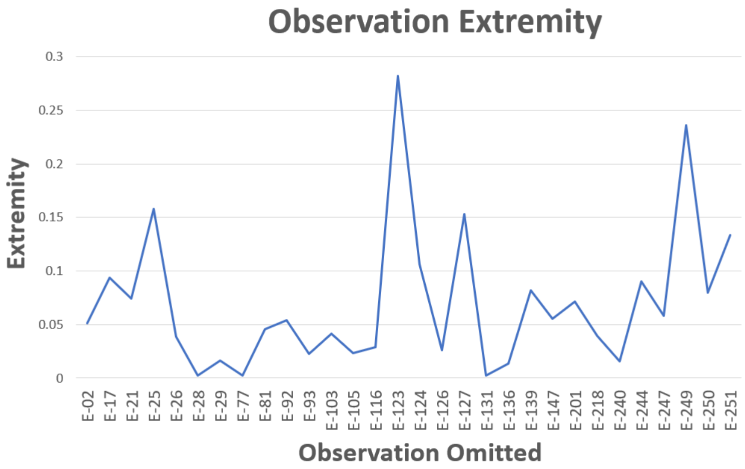


Figure 5.6: Observation Extremity: Visualization of Table 5.3 indicates the wide differences in effect of removing a observation's EoE data in training a model.

0.946 in comparison to the least optimal factor combination which has a desirability of 0.168.

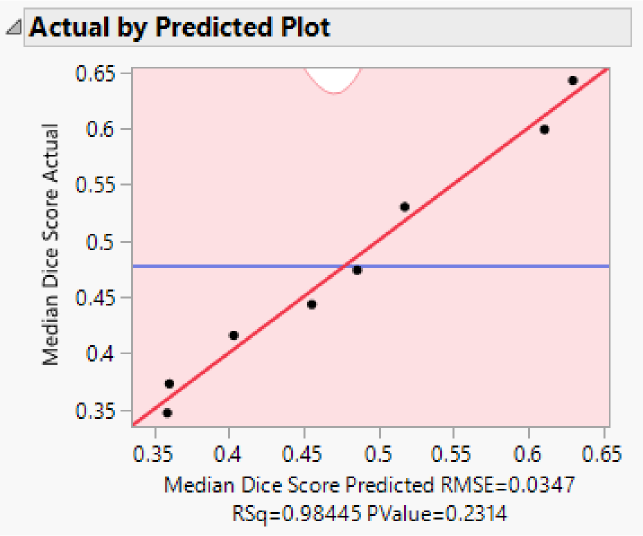
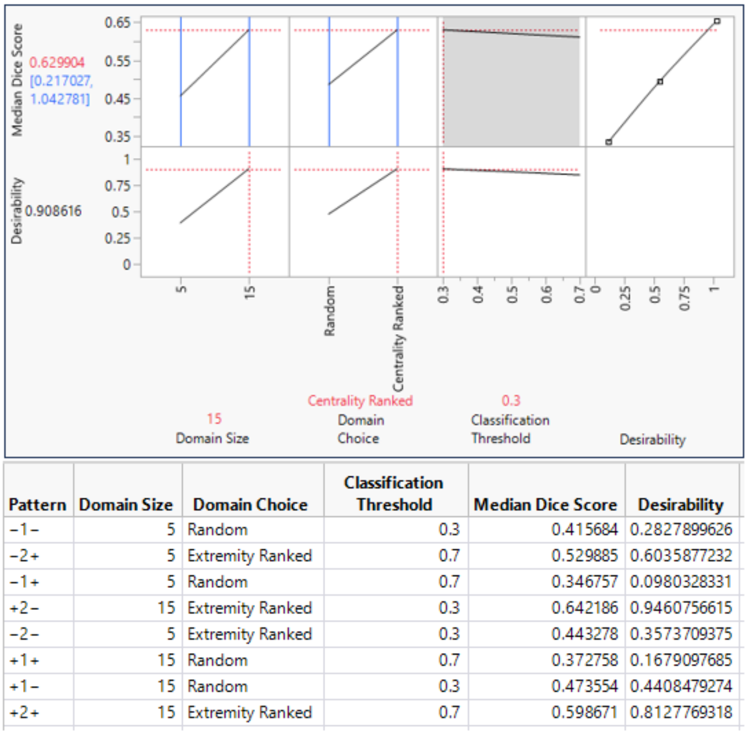


Figure 5.7: JMP output showing Standard Least Squares Plot to assess Effect Screening. Strong R^2 value of 0.9845 indicates the factors are appropriate choices but the p-value of 0.2314 is relatively high. This can be effected by the large variance created by the relatively small dataset.



Pattern	Domain Size	Domain Choice	Classification Threshold	Median Dice Score	Desirability
−1−	5	Random	0.3	0.415684	0.2827899626
−2+	5	Extremity Ranked	0.7	0.529885	0.6035877232
−1+	5	Random	0.7	0.346757	0.0980328331
+2−	15	Extremity Ranked	0.3	0.642186	0.9460756615
−2−	5	Extremity Ranked	0.3	0.443278	0.3573709375
+1+	15	Random	0.7	0.372758	0.1679097685
+1−	15	Random	0.3	0.473554	0.4408479274
+2+	15	Extremity Ranked	0.7	0.598671	0.8127769318

Figure 5.8: JMP output showing Full Factorial Experimental Design Setup and Optimal Factor Values: The optimal values for maximizing a multisource domain adversarial network dice score are Domain Size: 15, Domain Choice: Extremity Ranked, and Classification Threshold of 0.3.

Chapter 6

Conclusions and Future Work

In this dissertation, deep learning methods were first evaluated on two different biomedical image segmentation projects. The first dataset focusing on Brain Image Segmentation represented a strong baseline dataset[40] in which many current deep learning approaches are evaluated on. The results from using a MCD UNet on the the Brain Image Segmentation dataset justified further approaches in using the UNet on future medical image datasets. Furthermore, the MCD UNet provided a valuable visualization of the entropy present in prediction results. This research shows that this entropy is highly sensitive to data perturbations. A direct example of this is through observation E-139 where the lack of perturbations led to this observation having the largest negative impact on the model's entropy. Although the MCD UNet did not directly outperform Adorno et al's[2] approach, the MCD UNet median dice score of 0.591 was well within the range of the median dice scores Adorno et al observed which ranged from 0.517 - 0.665, illustrating comparable results. Additionally, this UNet provided an entropy visualization through the Monte Carlo Dropout. This introduced the second dataset with Eosinophilic Esophagitis taken from UVA Medical Center. To verify applicability of the UNet on this medical dataset, this Monte Carlo Dropout UNet segmentation dice score was compared against other segmentation approaches including Adorno et al[2]. With the current rise in Diffusion and Domain Adaptation as appropriate approaches for improving segmentation performance, a Denoising Diffusion Probabilistic Model (DDPM)[36] and a Multi-Domain Adversarial Network (MDAN)[105] performance were compared on the EoE dataset by treating all observations but one as the source domain and then using the remaining observation as the target domain. Results showed that the MDAN approach showed comparable performance to the DDPM while allowing more flexibility in how to conduct deep learning training. Furthermore, with the medical field firmly rooted in “patient” based results and placing “patients first” in care, the domain adaptation approach allows a deep learning method to focus on specific patients or groups of patients to assist analysis. Finally, the extremity ranked domain selection (ERDS) approach combines this MDAN approach with a extremity metric to assess the effect of removing a patient’s observations or groups of patients’ observations through a full factorial statistical design. ERDS provides strong evidence that the training data should be adequately analyzed before performing any serious deep learning approaches. This work also gives motivation towards further work in domain adaptation due to the incorporation of observation-centric analysis.

This research focuses specifically on an EoE dataset with source and target domains taken from the same dataset. One of the most important aspects of domain adaptation is ensuring that the source and target domains are different in order to test a model’s performance. This research fulfills this requirement since patient characteristics such as age, sex, ethnicity, BMI, location of biopsy taken, and time of initial biopsy collection are not the same across all observations. Thus, even though the source and target domains come from the same dataset for the same medical condition, this research provides valuable information as to the different ways EoE presents in patients and provides insight into the different disease phenotypes present. This research is disease specific and only focuses on EoE. As a possible future work, expanding on this approach and using other datasets for other medical conditions can provide information as to how segmentation models trained on EoE observations can help inform other disease phenotypes.

By having observation-based results drive a subsequent approach, this deep learning method parallels how actual medical providers care for patients. Finally, combining a traditional statistical approach with a new deep learning method demonstrates that although deep learning is moving extremely quickly as a field, some traditional techniques may have use and can produce powerful results. The full factorial design demonstrated that optimally selecting the factors of domain size (5, 15), domain choice (Random, Extremity Ranked), and classification threshold (0.3, 0.7) can significantly improve the dice score. For comparison, the baseline MCD UNet had a median dice score of 0.576 while the MDAN approach with optimal factors selected had a median dice score of 0.642.

As stated in the introduction, this dissertation contributed to many fields including data science, image segmentation, machine learning, deep learning, computer vision, artificial intelligence, medical image analysis, domain adaptation, and more. For clarity, the major contributions are repeated below:

- Demonstrated through experimentation that Bayesian optimization, specifically through Monte-Carlo Dropout, can produce entropy visualizations per patch that provide insight into the abilities of deep segmentation approaches and into the deep phenotyping present in medical data observations. The evaluation of these experimentations through a Dice Score directly addresses some of the issues Kobayashi saw in CVPR 2023 regarding using typical approaches such as Binary Cross Entropy on imbalanced problems[50]. Additionally, the entropy metric represents the variability of the predicted labels, giving medical researchers more information about areas of interest and medical image segmentation detection capabilities.

- Evaluated through experimentation that Multi-Domain methods have comparable performance to Diffusion based methods while providing more control over model training. This control allows observation-level analysis in determining the effect each observation has on deep segmentation model results. The motivation centers on modern approaches that have shown GAN approaches performing extremely effectively for domain adaptation methods[107][99][49]. Additionally, GAN approaches can address domain shift by using a generator to project features to an image space and a discriminator operates on this projected space[89][74]. With multidomain methods giving researchers more control over the model, a multidomain approach with GANs can leverage the strong performance of GANs with domain adaptation and allow observation-level analysis.

- Created a new metric called patient observation extremity (POE) for evaluating deep learning approaches by dropping subsets of training data and evaluating the effects of these omissions. This metric is a direct representation of a deep learning model's reliance on a subset of a dataset and can provide valuable information about what model performance will be when certain features of a dataset are not present. Based on literature review, the Leave-One-Out approaches in deep learning are limited to only cross validation in order to evaluate models[43]. In this research, the Leave-One Out approach focuses on observation-level differences in the training data that are not captured in standard deep learning approaches. A limitation of this Leave-One-Out approach for training is that the computational cost increases significantly when the training dataset increases. Given N observations, the model has to be re-trained N times in order to obtain the POE metric for each observation. This cost becomes even more of a factor if a model with more parameters is used.

- Showed the relationship of domain size, domain choice, and threshold value in domain adaptation. Full factorial designs encompass each pairing of these parameters driving efforts to improve future domain adaptation techniques. Furthermore, segmentation results vary between different source/target domain pairings. Intentionally setting source and target domains based off of the impact of each observation improves deep segmentation results. This contribution mirrors what Senhaji, et al observed in 2021 with their adaptive multidomain approach. Namely, choosing domains effectively in multidomain approaches can have significant impact on results due to the different characteristics represented in the domains[77].

- Produced optimal domain adaptation parameters that verified the importance of the new patient observation extremity (POE) metric on deep learning models. Setting these parameters significantly increased a multidomain adaptation method's dice score, demonstrating that the model is able to better identify areas of interest in an input image. Accompanying observation-level entropy quantification results provide information into how each observation contributes to model results. This directly addresses the main issue and problem for this dissertation - the lack of an observation-level approach for tuning segmentation models. This research creates a proven approach for future medical image segmentation: focus on the various features represented by the observations in the training data and use these features to create effective domains that will inform decisions.

- Used a traditional statistical approach through a full factorial design to evaluate a multi-domain adaptation approach created by the new extremity metric. Bonding a traditional approach with a modern one and achieving strong results indicates that future work should consider using traditional statistical and mathematical methods to support current approaches. Deep learning methods tend to focus on optimizing architectures[45], minimizing the reliance on labeled data[83], and better detection [62][61]. Future work should supplement these approaches with traditional statistical methods to analyze observation-level factor contributions.

- Improved the detection performance of two different biomedical image segmentation projects that all have critical clinical importance. The current field of medical image segmentation has a wealth of research towards cellular detection and identification [38][68][94][30][58]. Furthering these efforts allow medical researchers to gain more insight into how each of these conditions present in different datasets.

Discussion with medical researchers at the UVA GI Data Science lab has revealed a few possible future approaches for work. First, the multi-domain adaptation can be expanded to address imbalances in the amount of data observations have. The number of cells in the patches in the EoE dataset vary significantly with some patches containing cells in over half of all pixels in the patch to other patches with large areas with no cells at all. Given that training a model to identify cells performs better when there is more information about what the cells looks like and also what other cells around the interested cells look like, the patches can be placed into high, medium, and low cell densities for domains and deep learning models can be evaluated on them. Furthermore, there is not a specific size at which a biopsy is deemed insufficient and to safeguard against not being able to make a diagnosis, clinicians collect multiple biopsies from each region of concern. In the case of EoE, clinicians collect multiple biopsies from each region of the esophagus for patients who are being investigated for EoE, thereby increasing the likelihood that some of the samples will be sufficient. This led to some patient samples having only 10 labeled patches in the dataset while some other patient samples had over 70. This dissertation focused on extremity but using the number of patches as possible way of defining different domains in deep learning approaches can provide insight into if observations with larger amount of patches significantly affect dice scores or other measures of segmentation performance. Finally, one of the most significant limitations of this work is that the EoE dataset only has one person of color represented with the rest of the 29 patients being non-Hispanic white people. Future work should immediately strive to find data that encompasses a wider amount of diversity to ensure that effective deep phenotyping is not only driven by people of one race or one ethnicity.

Bibliography

[1] Shahira Abousamra et al. "Topology-Guided Multi-Class Cell Context Generation for Digital Pathology". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2023), pp. 3323–3333. DOI: 10.1109/cvpr52729.2023.00324.

[2] William Adorno et al. "Advancing eosinophilic esophagitis diagnosis and phenotype assessment with deep learning computer vision". In: *BIOIMAGING 2021 - 8th International Conference on Bioimaging; Part of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2021* (2021), pp. 44–55. DOI: 10.5220/0010241900440055.

[3] Mohammed Alhamid. *What is Cross-Validation?* 2020. URL: https://towardsdatascience.com/what-is-cross-validation-60c01f9d9e75.

[4] Ryad Amdoun et al. "The Desirability Optimization Methodology; a Tool to Predict Two Antagonist Responses in Biotechnological Systems: Case of Biomass Growth and Hyoscyamine Content in Elicited Datura starmonium Hairy Roots". In: *Iranian Journal of Biotechnology* 16.1 (2018), pp. 11–19. DOI: 10.21859/ijb.1339.

[5] Ujjwal Baid et al. "The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification". In: *eprint arXiv:2107.02314* (2021). URL: http://arxiv.org/abs/2107.02314.

[6] Spyridon Bakas et al. "Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival

Prediction in the BRATS Challenge”. In: *eprint arXiv:1811.02629* (2018). URL: http://arxiv.org/abs/1811.02629.

[7] Shai Ben-David et al. “A theory of learning from different domains”. In: *Machine Learning* 79.1-2 (2010), pp. 151–175. ISSN: 15730565. DOI: 10.1007/s10994-009-5152-4.

[8] Shai Ben-David et al. “Analysis of Representations for Domain Adaptation”. In: *NIPS 2006: Proceedings of the 19th International Conference on Neural Information Processing Systems* (2006), pp. 137–144. DOI: 10.7551/mitpress/7503.003.0022.

[9] Stuart Carr, Edmond S. Chan, and Wade Watson. “Eosinophilic esophagitis”. In: *Allergy, Asthma and Clinical Immunology* 14.s2 (2018), pp. 1–11. ISSN: 17101492. DOI: 10.1186/s13223-018-0287-0. URL: https://doi.org/10.1186/s13223-018-0287-0.

[10] Tsung Han Chan et al. “PCANet: A Simple Deep Learning Baseline for Image Classification?” In: *IEEE Transactions on Image Processing* 24.12 (2015), pp. 5017–5032. ISSN: 10577149. DOI: 10.1109/TIP.2015.2475625.

[11] Hanzhi Chen et al. “TexPose: Neural Texture Learning for Self-Supervised 6D Object Pose Estimation”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2023-June (2023), pp. 4841–4852. ISSN: 10636919. DOI: 10.1109/CVPR52729.2023.00469.

[12] Zhaozheng Chen et al. “Class Re-Activation Maps for Weakly-Supervised Semantic Segmentation”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2022-June (2022), pp. 959–968. ISSN: 10636919. DOI: 10.1109/CVPR52688.2022.00104.

[13] Ulysse Côté-Allard et al. "Deep Learning for Electromyographic Hand Gesture Signal Classification Using Transfer Learning". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 27.4 (2019), pp. 760–771. ISSN: 15344320. DOI: 10.1109/TNSRE.2019.2896269.

[14] Tomer Czyzewski et al. "Machine Learning Approach for Biopsy-Based Identification of Eosinophilic Esophagitis Reveals Importance of Global features". In: *IEEE Open Journal of Engineering in Medicine and Biology* 2 (2021), pp. 218–223. ISSN: 26441276. DOI: 10.1109/OJEMB.2021.3089552.

[15] Nati Daniel et al. "PECNet : A Deep Multi-Label Segmentation Network for Eosinophilic Esophagitis Biopsy Diagnostics". In: *eprint arXIV:2103.02015* (2021), pp. 1–12.

[16] Evan S. Dellon. "Epidemiology of eosinophilic esophagitis". In: *Gastroenterology Clinics of North America* 43.2 (2014), pp. 201–218. ISSN: 15581942. DOI: 10.1016/j.gtc.2014.02.002.

[17] Roberto Dessì et al. "Cross-Domain Image Captioning with Discriminative Finetuning". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2023-June (2023), pp. 6935–6944. ISSN: 10636919. DOI: 10.1109/CVPR52729.2023.00670.

[18] Terrance DeVries and Graham W. Taylor. "Leveraging Uncertainty Estimates for Predicting Segmentation Quality". In: *Proceedings for the 1st Conference on Medical Imaging with Deep Learning (MIDL 2018), Amsterdam, The Netherlands.* (2018). URL: http://arxiv.org/abs/1807.00502.

[19] Lee Dice. "Measures of the Amount of Ecologic Association Between Species". In: *Ecology* 26.3 (1945), pp. 297–302. DOI: 10.2307/1932409.

[20] Abdulkerim Duman et al. “RFS+: A Clinically Adaptable and Computationally Efficient Strategy for Enhanced Brain Tumor Segmentation”. In: *Cancers* 15.23 (2023), pp. 1–21. ISSN: 20726694. DOI: 10.3390/cancers15235620.

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

[22] Chao Feng, Ziyang Chen, and Andrew Owens. “Self-Supervised Video Forensics by Audio-Visual Anomaly Detection”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2023), pp. 10491–10503. DOI: 10.1109/cvpr52729.2023.01011.

[23] James P. Franciosi et al. “Proton Pump Inhibitor Therapy for Eosinophilic Esophagitis: History, Mechanisms, Efficacy, and Future Directions”. In: *Journal of Asthma and Allergy* 15.January (2022), pp. 281–302. ISSN: 11786965. DOI: 10.2147/JAA.S274524.

[24] Glenn T. Furuta et al. “Eosinophilic Esophagitis in Children and Adults: A Systematic Review and Consensus Recommendations for Diagnosis and Treatment. Sponsored by the American Gastroenterological Association (AGA) Institute and North American Society of Pediatric Gastroenterol”. In: *Gastroenterology* 133.4 (2007), pp. 1342–1363. ISSN: 00165085. DOI: 10.1053/j.gastro.2007.08.017.

[25] Michał Futrega et al. “Optimized U-Net for Brain Tumor Segmentation”. In: *arXiv preprint arXiv:2110.03352* (2021), pp. 1–15. URL: http://arxiv.org/abs/2110.03352.

[26] Pius Kwao Gadosey et al. "SD-UNET: Stripping down U-net for segmentation of biomedical images on platforms with low computational budgets". In: *Diagnostics* 10.2 (2020). ISSN: 20754418. DOI: `10.3390/diagnostics10020110`.

[27] Yarin Gal and Zoubin Ghahramani. "Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference". In: *Proceedings of ICLR 2016* (2015), pp. 1–12. URL: `http://arxiv.org/abs/1506.02158`.

[28] Yarin Gal and Zoubin Ghahramani. "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning". In: *33rd International Conference on Machine Learning, ICML 2016* 3 (2016), pp. 1651–1660.

[29] Yaroslav Ganin et al. "Domain-Adversarial Training of Neural Networks". In: *Journal of Machine Learning Research* 17 (2016), pp. 1–35.

[30] Mohsen Ghafoorian et al. "Transfer learning for domain adaptation in MRI: Application in brain lesion segmentation". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 10435 LNCS (2017), pp. 516–524. ISSN: 16113349. DOI: `10.1007/978-3-319-66179-7{\_}59`.

[31] Hao Guan and Mingxia Liu. "Domain Adaptation for Medical Image Analysis: A Survey". In: *IEEE Transactions on Biomedical Engineering* 69.3 (2022), pp. 1173–1185. ISSN: 15582531. DOI: `10.1109/TBME.2021.3117407`.

[32] Evan Hann et al. "Ensemble of Deep Convolutional Neural Networks with Monte Carlo Dropout Sampling for Automated Image Segmentation Quality Control and Robust Deep Learning Using Small Datasets". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12722 LNCS.July (2021), pp. 280–293. ISSN: 16113349. DOI: `10.1007/978-3-030-80432-9{\_}22`.

[33] Mehdi Hassan et al. “Developing intelligent medical image modality classification system using deep transfer learning and LDA”. In: *Scientific Reports* 10.1 (2020), pp. 1–14. ISSN: 20452322. DOI: `10.1038/s41598-020-69813-2`. URL: `https://doi.org/10.1038/s41598-020-69813-2`.

[34] Haibo He and Yunqian Ma. *Imbalanced Learning: Foundations, Algorithms, and Applications*. The Institute of Electrical and Electronics Engineers, Inc., 2013, p. 188. ISBN: 9781118074626. DOI: `10.1002/9781118646106`.

[35] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2016-Decem (2016), pp. 770–778. ISSN: 10636919. DOI: `10.1109/CVPR.2016.90`.

[36] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising diffusion probabilistic models”. In: *Advances in Neural Information Processing Systems* 2020-Decem.NeurIPS 2020 (2020), pp. 1–25. ISSN: 10495258.

[37] Tzu Ming Harry Hsu et al. “Unsupervised domain adaptation with imbalanced cross-domain data”. In: *Proceedings of the IEEE International Conference on Computer Vision* 2015 Inter (2015), pp. 4121–4129. ISSN: 15505499. DOI: `10.1109/ICCV.2015.469`.

[38] Gao Huang et al. “Densely connected convolutional networks”. In: *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* 2017-Janua (2017), pp. 2261–2269. DOI: `10.1109/CVPR.2017.243`.

[39] Huimin Huang et al. “WNET : AN END-TO-END ATLAS-GUIDED AND BOUNDARY-ENHANCED NETWORK FOR MEDICAL IMAGE SEGMENTATION”. In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)* (2020), pp. 763–766.

[40] Yawen Huang et al. "Brain Image Synthesis with Unsupervised Multivariate Canonical CSC 4Net". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2021), pp. 5877–5886. ISSN: 10636919. DOI: 10.1109/CVPR46437.2021.00582.

[41] Fabian Isensee et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature Methods* 18.2 (2021), pp. 203–211. ISSN: 15487105. DOI: 10.1038/s41592-020-01008-z. URL: http://dx.doi.org/10.1038/s41592-020-01008-z.

[42] Junbong Jang, Kwonmoo Lee, and Tae Kyun Kim. "Unsupervised Contour Tracking of Live Cells by Mechanical and Cycle Consistency Losses". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2023-June (2023), pp. 227–236. ISSN: 10636919. DOI: 10.1109/CVPR52729.2023.00030.

[43] Ruoxi Jia et al. "Scalability vs. Utility: Do We Have to Sacrifice One for the Other in Data Importance Quantification?" In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2021), pp. 8235–8243. ISSN: 10636919. DOI: 10.1109/CVPR46437.2021.00814.

[44] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding". In: *British Machine Vision Conference 2017, BMVC 2017* (2017). DOI: 10.5244/c.31.57.

[45] Asifullah Khan et al. "A survey of the recent architectures of deep convolutional neural networks". In: *Artificial Intelligence Review* 53.8 (2020), pp. 5455–5516. ISSN: 15737462. DOI: 10.1007/s10462-020-09825-6.

[46] Hyeonseong Kim et al. “Single Domain Generalization for LiDAR Semantic Segmentation”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2023-June (2023), pp. 17587–17598. issn: 10636919. doi: `10.1109/CVPR52729.2023.01687`.

[47] Jaeill Kim et al. “VNE: An Effective Method for Improving Deep Representation by Manipulating Eigenvalue Distribution”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2023-June (2023), pp. 3799–3810. issn: 10636919. doi: `10.1109/CVPR52729.2023.00370`.

[48] Mijung Kim, Jasper Zuallaert, and Wesley De Neve. “Few-shot learning using a small-sized dataset of high-resolution FUNDUS images for glaucoma diagnosis”. In: *MMHealth 2017 - Proceedings of the 2nd International Workshop on Multimedia for Personal Health and Health Care, co-located with MM 2017* (2017), pp. 89–92. doi: `10.1145/3132635.3132650`.

[49] Taeksoo Kim et al. “Learning to discover cross-domain relations with generative adversarial networks”. In: *34th International Conference on Machine Learning, ICML 2017* 4 (2017), pp. 2941–2949.

[50] Takumi Kobayashi. “Two-way Multi-Label Loss”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2023), pp. 7476–7485.

[51] Simon A.A. Kohl et al. “A probabilistic U-net for segmentation of ambiguous images”. In: *Advances in Neural Information Processing Systems* 2018-Decem.NeurIPS (2018), pp. 6965–6975. issn: 10495258.

[52] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *Nature* 521.7553 (2015), pp. 436–444. issn: 14764687. doi: `10.1038/nature14539`.

[53] Haoliang Li et al. “Domain generalization for medical imaging classification with linear-dependency regularization”. In: *Advances in Neural Information Processing Systems* 2020-Decem.NeurIPS (2020), pp. 1–14. ISSN: 10495258.

[54] Kehan Li et al. “ACSeg: Adaptive Conceptualization for Unsupervised Semantic Segmentation”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2023-June (2023), pp. 7162–7172. ISSN: 10636919. DOI: 10.1109/CVPR52729.2023.00692.

[55] Mengfang Li et al. “Medical image analysis using deep learning algorithms”. In: *Frontiers in Public Health* 11.November (2023), pp. 1–28. ISSN: 22962565. DOI: 10.3389/fpubh.2023.1273253.

[56] Ruihuang Li et al. “DynaMask: Dynamic Mask Selection for Instance Segmentation”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2023-June (2023), pp. 11279–11288. ISSN: 10636919. DOI: 10.1109/CVPR52729.2023.01085.

[57] Kevin Lin et al. “Diffusion and Multi-Domain Adaptation Methods for Eosinophil Segmentation”. In: *ACM International Conference Proceeding Series* (2024), pp. 12–15. URL: http://arxiv.org/abs/2403.11323.

[58] Kevin Lin et al. “Uncertainty Quantification for Eosinophil Segmentation”. In: *ACM International Conference Proceeding Series* (2023), pp. 15–19. DOI: 10.1145/3632047.3632050.

[59] Yang Liu, Zhipeng Zhou, and Baigui Sun. “COT: Unsupervised Domain Adaptation with Clustering and Optimal Transport”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2023), pp. 19998–20007. DOI: 10.1109/cvpr52729.2023.01915.

[60] Zhanxin Ma et al. “Using Shannon Entropy to Improve the Identification of MP-SBM Models with Undesirable Output”. In: *Entropy* 24.11 (2022), pp. 1–20. ISSN: 10994300. DOI: 10.3390/e24111608.

[61] Nazanin Moradinasab et al. “Label-efficient Contrastive Learning-based model for nuclei detection and classification in 3D Cardiovascular Immunofluorescent Images”. In: *Workshop on Medical Image Learning with Limited and Noisy Data* (2023), pp. 1–11.

[62] Nazanin Moradinasab et al. “Weakly Supervised Deep Instance Nuclei Detection using Points Annotation in 3D Cardiovascular Immunofluorescent Images”. In: *Proceedings of Machine Learning Research Volumne 182* (2022), pp. 4–5. ISSN: 26403498. URL: http://arxiv.org/abs/2208.00098.

[63] Gustav Müller-Franzes et al. “Diffusion Probabilistic Models beat GANs on Medical Images”. In: *eprint arXiv:2212.07501* (2022). URL: http://arxiv.org/abs/2212.07501.

[64] Barack Obama. *The Precision Medicine Initiative.* 2015. URL: https://obamawhitehouse.archives.gov/precision-medicine.

[65] Sinno Jialin Pan and Qiang Yang. “A survey on transfer learning”. In: *IEEE Transactions on Knowledge and Data Engineering* 22.10 (2010), pp. 1345–1359. ISSN: 10414347. DOI: 10.1109/TKDE.2009.191.

[66] Himashi Peiris et al. “Reciprocal Adversarial Learning for Brain Tumor Segmentation: A Solution to BraTS Challenge 2021 Segmentation Task”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12962 LNCS (2022), pp. 171–181. ISSN: 16113349. DOI: 10.1007/978-3-031-08999-2{_}13.

[67] Suraj Regmi. *Entropy, Cross Entropy, and KL Divergence*. 2020. URL: https://towardsdatascience.com/entropy-cross-entropy-and-kl-divergence-17138ffab87b.

[68] Jacob C. Reinhold et al. “Validating Uncertainty in Medical Image Translation”. In: *Proceedings - International Symposium on Biomedical Imaging* 2020-April (2020), pp. 95–98. ISSN: 19458452. DOI: 10.1109/ISBI45749.2020.9098543.

[69] Chuan Xian Ren et al. “Transfer learning of structured representation for face recognition”. In: *IEEE Transactions on Image Processing* 23.12 (2014), pp. 5440–5454. ISSN: 10577149. DOI: 10.1109/TIP.2014.2365725.

[70] April Reynolds. “Patient-centered Care”. In: *Radiology Technology* 81.2 (2009), pp. 133–147. URL: https://pubmed.ncbi.nlm.nih.gov/19901351/.

[71] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *International Conference on Medical image computing and computer-assisted intervention* (May 2015), pp. 234–241. ISSN: 21693536. URL: http://arxiv.org/abs/1505.04597.

[72] Thomas M. Runge et al. “Causes and Outcomes of Esophageal Perforation in Eosinophilic Esophagitis”. In: *Journal of Clinical Gastroenterology* 51.9 (2017), pp. 805–813. ISSN: 15392031. DOI: 10.1097/MCG.0000000000000718.

[73] Parisa Saat et al. “A domain adaptation benchmark for T1-weighted brain magnetic resonance image segmentation”. In: *Frontiers in Neuroinformatics* 16 (2022). ISSN: 16625196. DOI: 10.3389/fninf.2022.919779.

[74] Swami Sankaranarayanan et al. “Learning from Synthetic Data : Addressing Domain Shift for Semantic Segmentation”. In: *Proceedings of the IEEE Com-*

puter Society Conference on Computer Vision and Pattern Recognition (2018), pp. 3752–3761. URL: http://openaccess.thecvf.com/content_cvpr_2018/papers/Sankaranarayanan_Learning_From_Synthetic_CVPR_2018_paper.pdf.

[75] Devvi Sarwinda et al. “Deep Learning in Image Classification using Residual Network (ResNet) Variants for Detection of Colorectal Cancer”. In: *Procedia Computer Science* 179.2019 (2021), pp. 423–431. ISSN: 18770509. DOI: 10.1016/j.procs.2021.01.025. URL: https://doi.org/10.1016/j.procs.2021.01.025.

[76] scikit-learn developers. *sklearn.model_selection.LeaveOneOut.* 2007. URL: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.LeaveOneOut.html.

[77] Ali Senhaji et al. “Not all domains are equally complex: Adaptive multi-domain learning”. In: *Proceedings - International Conference on Pattern Recognition* (2020), pp. 8663–8670. ISSN: 10514651. DOI: 10.1109/ICPR48806.2021.9412215.

[78] C. E. Shannon. “A Mathematical Theory of Communication”. In: *Bell System Technical Journal* 27.4 (1948), pp. 623–656. ISSN: 15387305. DOI: 10.1002/j.1538-7305.1948.tb00917.x.

[79] Hyungseob Shin et al. “SDC-UDA: Volumetric Unsupervised Domain Adaptation Framework for Slice-Direction Continuous Cross-Modality Medical Image Segmentation”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2023), pp. 7412–7421. DOI: 10.1109/cvpr52729.2023.00716.

[80] Jasdeep Singh, Subrahmanyam Murala, and G. Sankara Raju Kosuru. "Multi Domain Learning for Motion Magnification". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2023), pp. 13914–13923. DOI: `10.1109/cvpr52729.2023.01337`.

[81] Sandeep Singh, Benoy Singh, and Anuj Kumar. "Magnetic resonance imaging image-based segmentation of brain tumor using the modified transfer learning method". In: *Journal of Medical Physics* 47.4 (2022), pp. 315–321. ISSN: 19983913. DOI: `10.4103/jmp.jmp{\_}52{\_}22`.

[82] M. Stone. "Cross-Validatory Choice and Assessment of Statistical Predictions (With Discussion)". In: *Journal of the Royal Statistical Society* 38.1 (1976), p. 102. URL: `https://academic.oup.com/jrsssb/article/38/1/102/7027391`.

[83] Liyan Sun et al. "Few-shot medical image segmentation using a global correlation network with discriminative embedding". In: *Computers in Biology and Medicine* 140.November 2021 (2022), p. 105067. ISSN: 18790534. DOI: `10.1016/j.compbiomed.2021.105067`. URL: `https://doi.org/10.1016/j.compbiomed.2021.105067`.

[84] Shuhan Tan, Xingchao Peng, and Kate Saenko. "Class-Imbalanced Domain Adaptation: An Empirical Odyssey". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12535 LNCS (2020), pp. 585–602. ISSN: 16113349. DOI: `10.1007/978-3-030-66415-2{\_}38`.

[85] Edward C. Tribus Myron; McIrvine. "Energy and Information". In: *Scientific American* 225.3 (1971), pp. 179–190.

[86] Shashank Tripathi et al. "3D Human Pose Estimation via Intuitive Physics". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2023), pp. 4713–4725. DOI: 10.1109/cvpr52729.2023.00457.

[87] Vidit Vidit, Martin Engilberge, and Mathieu Salzmann. "CLIP the Gap: A Single Domain Generalization Approach for Object Detection". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2023-June.i (2023), pp. 3219–3229. ISSN: 10636919. DOI: 10.1109/CVPR52729.2023.00314.

[88] Fusheng Wang et al. "Managing and querying whole slide images". In: *Medical Imaging 2012: Advanced PACS-based Imaging Informatics and Therapeutic Applications* 8319 (2012), 83190J. ISSN: 0277786X. DOI: 10.1117/12.912388.

[89] Mei Wang and Weihong Deng. "Deep visual domain adaptation: A survey". In: *Neurocomputing* 312 (2018), pp. 135–153. ISSN: 18728286. DOI: 10.1016/j.neucom.2018.05.083.

[90] Yunyun Wang et al. "TIToK: A solution for bi-imbalanced unsupervised domain adaptation". In: *Neural Networks* 164 (2023), pp. 81–90. ISSN: 18792782. DOI: 10.1016/j.neunet.2023.04.027. URL: https://doi.org/10.1016/j.neunet.2023.04.027.

[91] Jason W. Wei et al. "Automated detection of celiac disease on duodenal biopsy slides: A deep learning approach". In: *Journal of Pathology Informatics* 10.1 (2019). ISSN: 21533539. DOI: 10.4103/jpi.jpi{_}87{_}18.

[92] Xide Xia and Brian Kulis. "W-Net: A Deep Model for Fully Unsupervised Image Segmentation". In: *eprint arXiv:1711.08506* (2017). URL: http://arxiv.org/abs/1711.08506.

[93] Ronald Xie et al. “MAESTER: Masked Autoencoder Guided Segmentation at Pixel Resolution for Accurate, Self-Supervised Subcellular Structure Recognition”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2023-June (2023), pp. 3292–3301. ISSN: 10636919. DOI: 10.1109/CVPR52729.2023.00321.

[94] Yan Xu et al. “Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features”. In: *BMC Bioinformatics* 18.1 (2017), pp. 1–17. ISSN: 14712105. DOI: 10.1186/s12859-017-1685-x.

[95] Wenjun Yan et al. “The Domain Shift Problem of Medical Image Segmentation and Vendor-Adaptation by Unet-GAN”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11765 LNCS (2019), pp. 623–631. ISSN: 16113349. DOI: 10.1007/978-3-030-32245-8{_}69.

[96] Antoine Yang et al. “Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2023-June (2023), pp. 10714–10726. ISSN: 10636919. DOI: 10.1109/CVPR52729.2023.01032.

[97] Jinyu Yang et al. “Resource-Efficient RGBD Aerial Tracking”. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (2023), pp. 13374–13383. ISSN: 21607516. DOI: 10.1109/CVPR52729.2023.01285.

[98] Jiayuan Ye et al. “Leave-one-out Distinguishability in Machine Learning”. In: (2023), pp. 1–37. URL: http://arxiv.org/abs/2309.17310.

[99] Zili Yi et al. “DualGAN: Unsupervised Dual Learning for Image-to-Image Translation”. In: *Proceedings of the IEEE International Conference on Computer Vision* 2017-Octob (2017), pp. 2868–2876. ISSN: 15505499. DOI: `10.1109/ICCV.2017.310`.

[100] Tao Zeng, Bian Wu, and Shuiwang Ji. “DeepEM3D: Approaching human-level performance on 3D anisotropic em image segmentation”. In: *Bioinformatics* 33.16 (2017), pp. 2555–2562. ISSN: 14602059. DOI: `10.1093/bioinformatics/btx188`.

[101] Guying Zhang et al. “Deep fusion of multi-modal features for brain tumor image segmentation”. In: *Heliyon* 9.8 (2023), e19266. ISSN: 24058440. DOI: `10.1016/j.heliyon.2023.e19266`. URL: `https://doi.org/10.1016/j.heliyon.2023.e19266`.

[102] Yunzhi Zhang et al. “Seeing a Rose in Five Thousand Ways ”5000” Generated Roses”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2023), pp. 962–971. URL: `https://cs.stanford.edu/`.

[103] Zhilu Zhang and Mert R. Sabuncu. “Generalized cross entropy loss for training deep neural networks with noisy labels”. In: *Advances in Neural Information Processing Systems* 2018-Decem.NeurIPS (2018), pp. 8778–8788. ISSN: 10495258.

[104] Han Zhao et al. “Multiple source domain adaptation with adversarial learning”. In: *6th International Conference on Learning Representations, ICLR 2018 - Workshop Track Proceedings* 2012 (2018), pp. 1–24.

[105] Han Zhao et al. "Multiple Source Domain Adaptation with Adversarial Training of Neural Networks". In: *NeurIPS 2018 Proceedings* (2017), pp. 1–24. URL: `http://arxiv.org/abs/1705.09684`.

[106] Donghao Zhou et al. "RepMode: Learning to Re-Parameterize Diverse Experts for Subcellular Structure Prediction". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2023), pp. 3312–3322. DOI: `10.1109/cvpr52729.2023.00323`.

[107] Jun Yan Zhu et al. "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks". In: *Proceedings of the IEEE International Conference on Computer Vision* 2017-Octob (2017), pp. 2242–2251. ISSN: 15505499. DOI: `10.1109/ICCV.2017.244`.

[108] Orr Zohar, Kuan Chieh Wang, and Serena Yeung. "PROB: Probabilistic Objectness for Open World Object Detection". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2023-June (2023), pp. 11444–11453. ISSN: 10636919. DOI: `10.1109/CVPR52729.2023.01101`.

Appendices

Appendix A

Observation Images

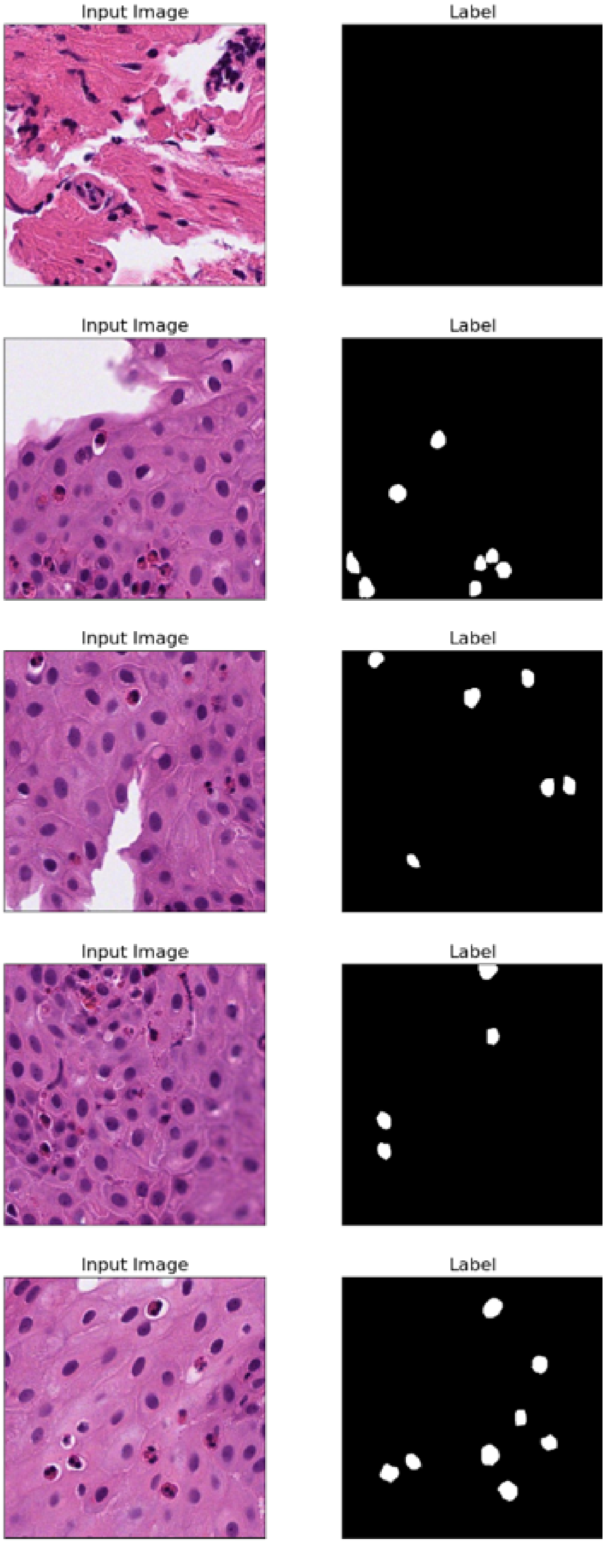


Figure A.1: Observation E-123 Images 1 of 2

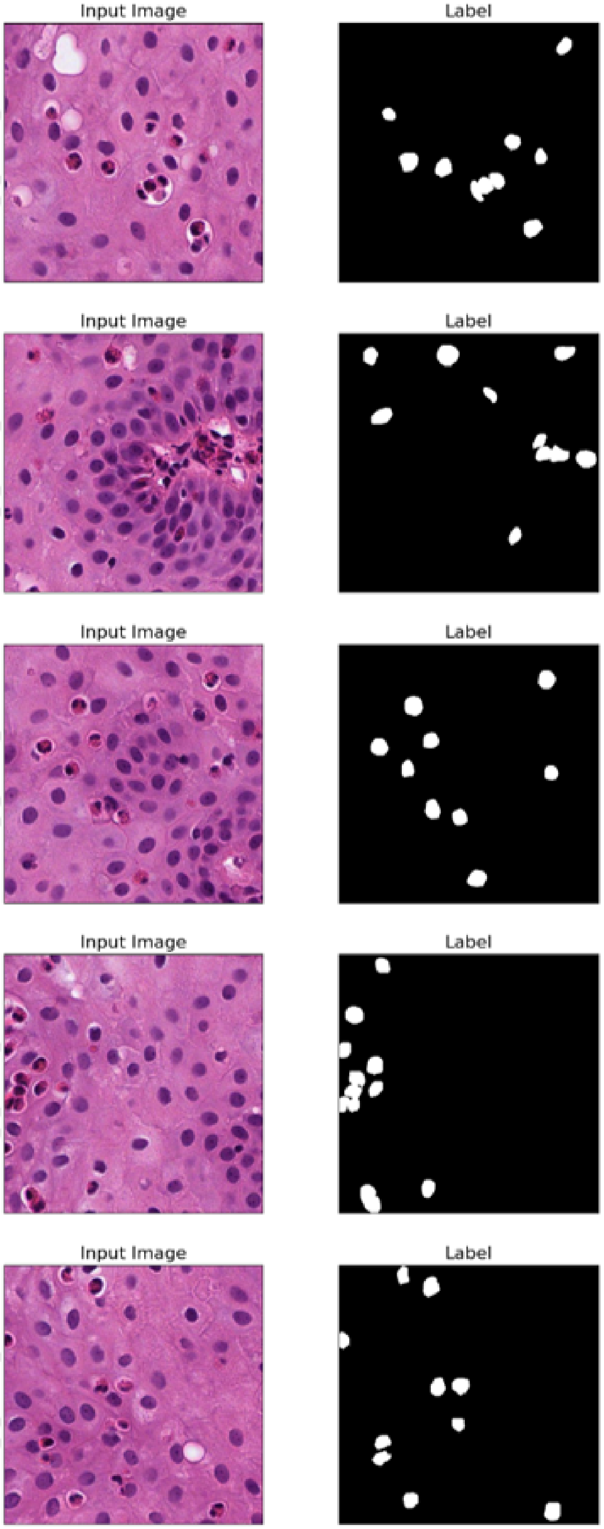


Figure A.2: Observation E-123 Images 2 of 2

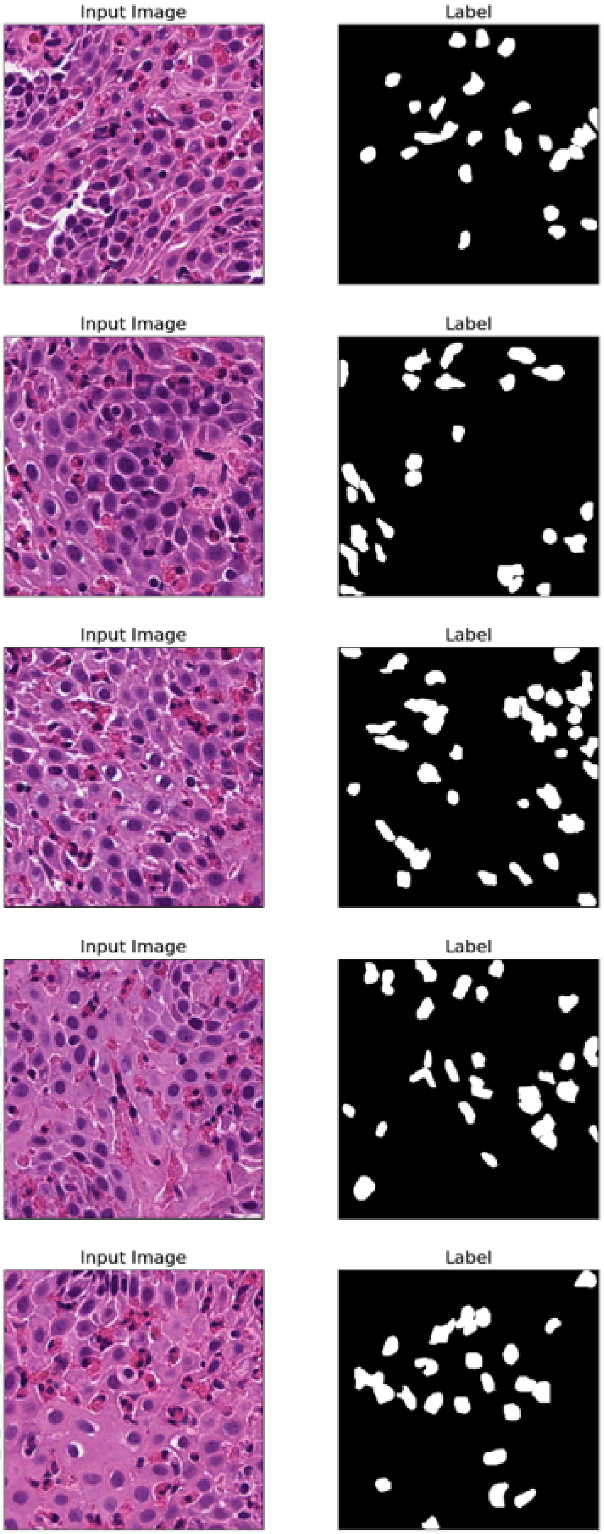


Figure A.3: Observation E-77 Images 1 of 2

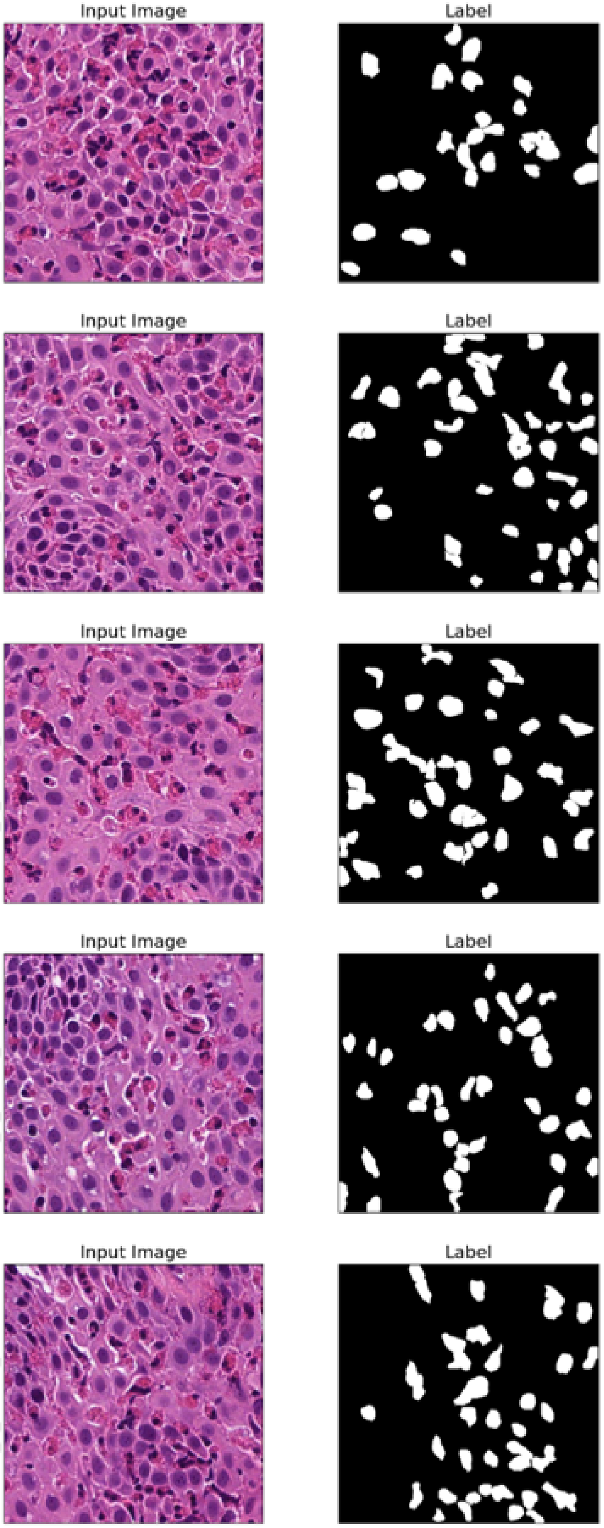


Figure A.4: Observation E-77 Images 2 of 2

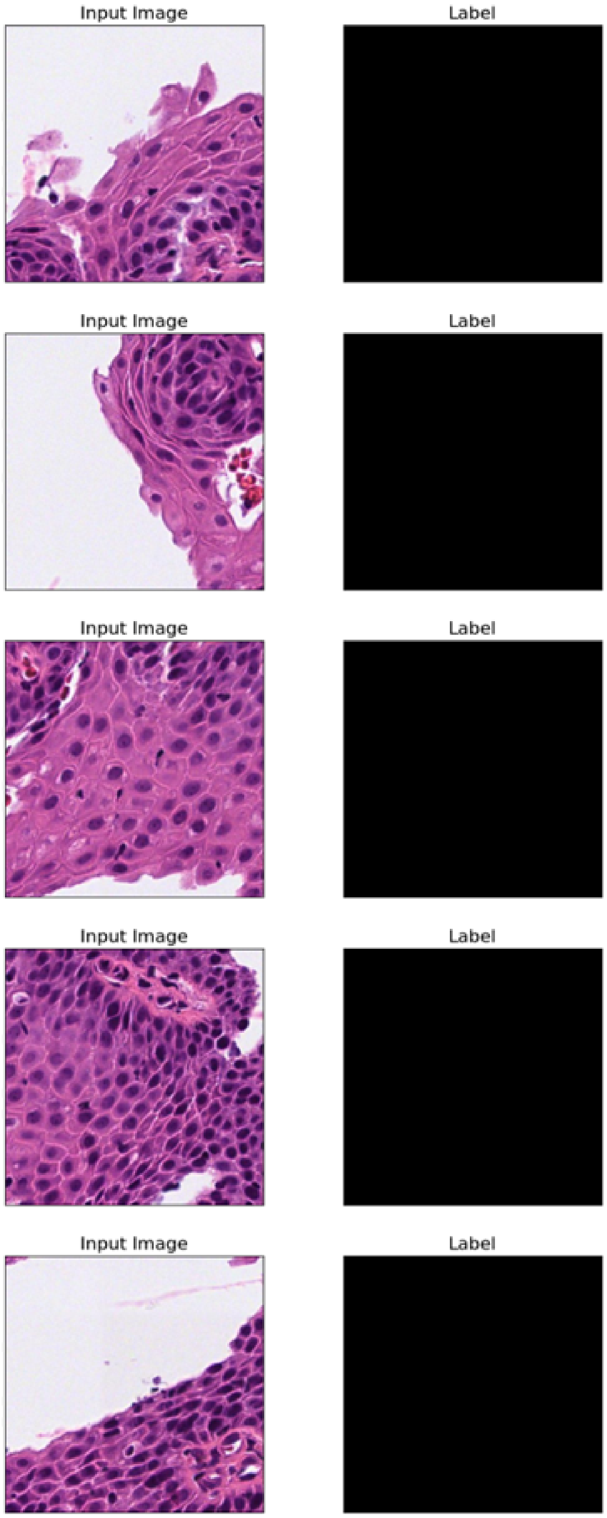


Figure A.5: Observation E-139 Images 1 of 4

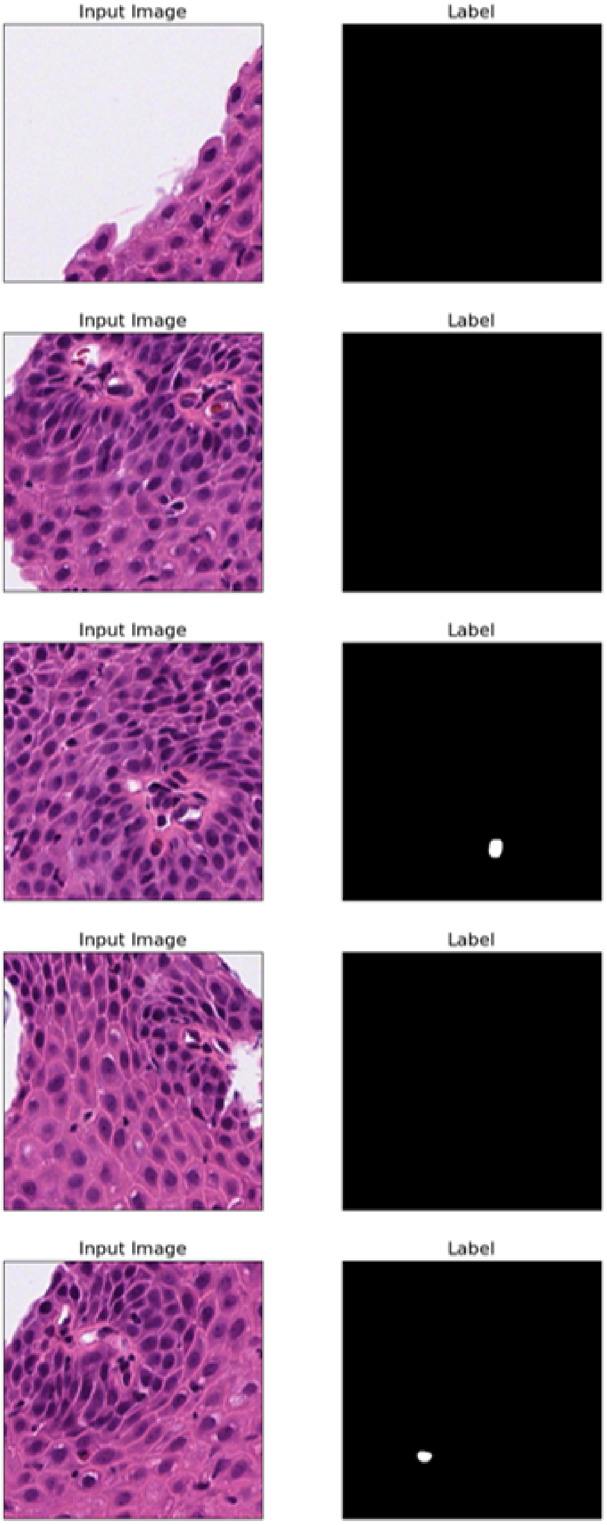


Figure A.6: Observation E-139 Images 2 of 4

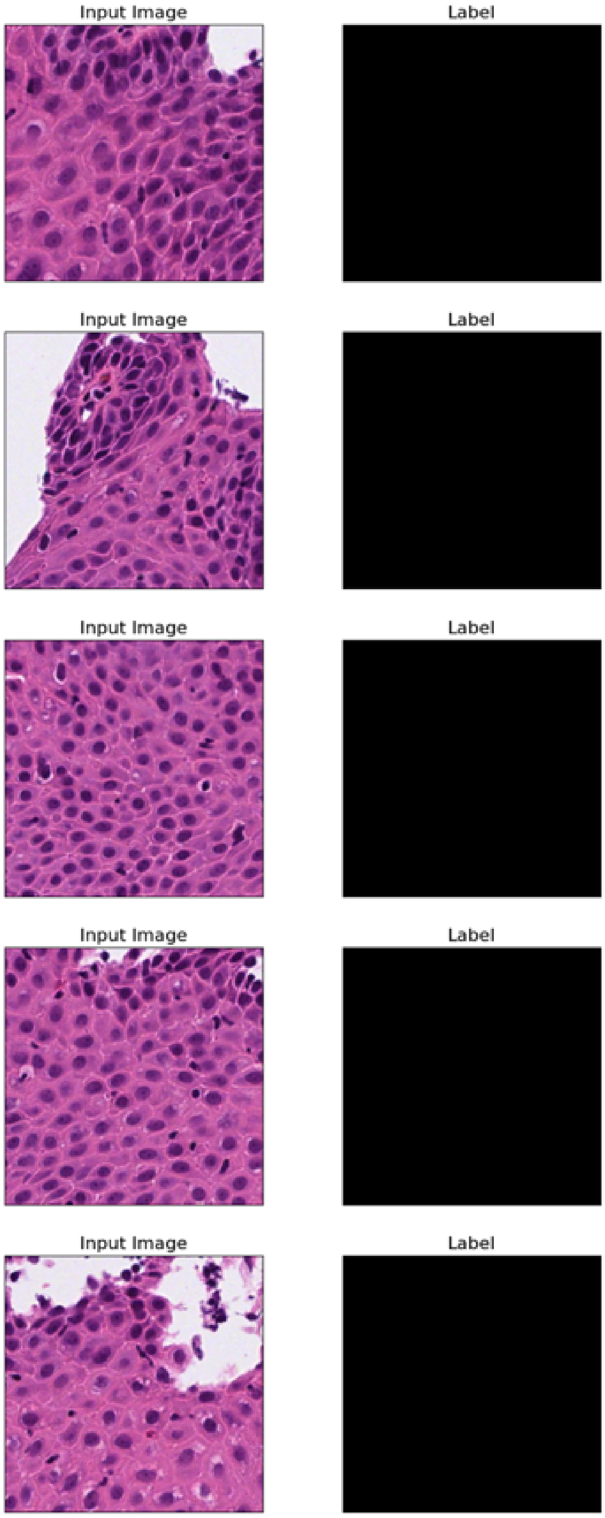


Figure A.7: Observation E-139 Images 3 of 4

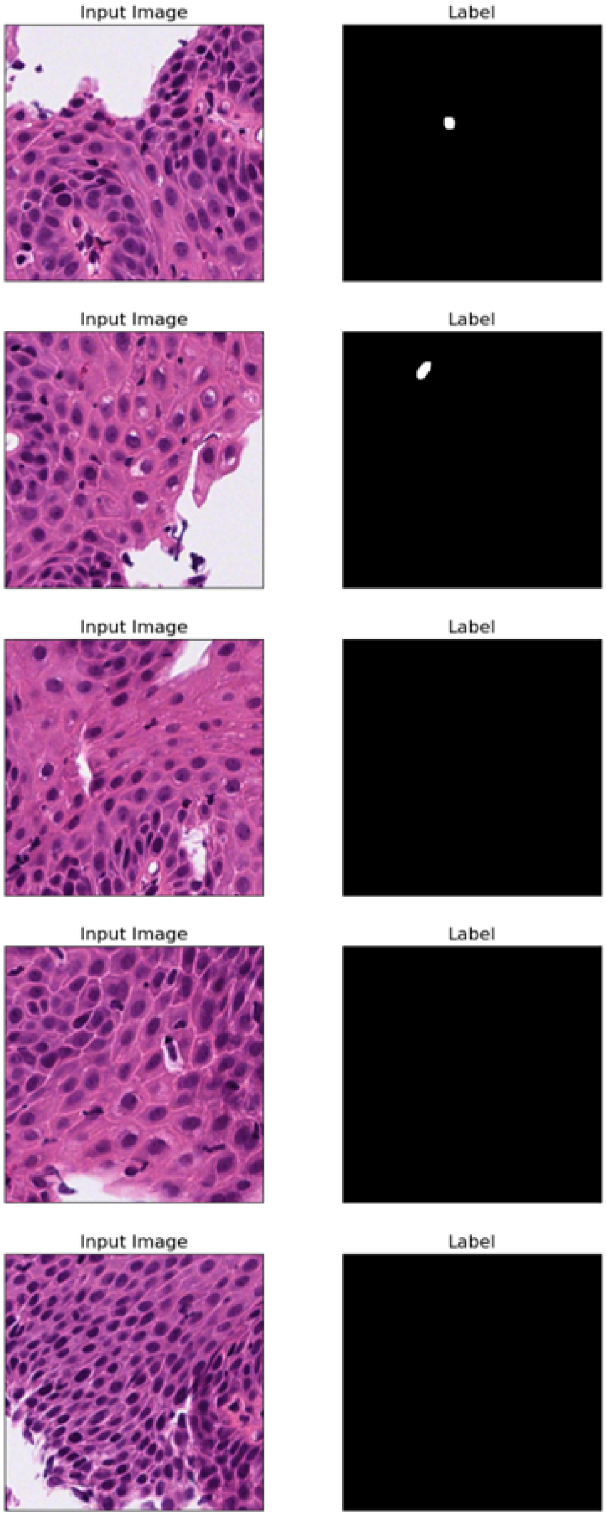


Figure A.8: Observation E-139 Images 4 of 4