APPROVAL SHEET

Title of Thesis: Reducing Bias in Tuberculosis Screening with Deep Domain Adaptation

Name of Candidate: Nishanjan Ravin
Master of Science, 2021

Thesis and Abstract Approved:

Dr. David Chapman
Assistant Professor
Department of Computer Science
and Electrical Engineering

Date Approved: 07/30/2021

ABSTRACT

Title of Thesis: Reducing Bias in Tuberculosis Screening with Deep Domain Adaptation

Nishanjan Ravin
Master of Science, 2021

Thesis directed by: Dr. David Chapman
Assistant Professor
Department of Computer Science
and Electrical Engineering

We demonstrate that Domain Invariant Feature Learning (DIFL) can improve the out-of-domain generalizability of a deep Tuberculosis screening algorithm.

It is well known that state of the art deep learning algorithms often have difficulty generalizing to unseen data distributions due to "domain shift". In the context of medical imaging, this could lead to unintended biases such as the inability to generalize from one patient population to another. We analyze the performance of a ResNet-50 classifier for the purposes of Tuberculosis screening using the four most popular public datasets with geographically diverse sources of imagery. We show that without domain adaptation, ResNet-50 has difficulty in generalizing between imaging distributions from a number of public Tuberculosis screening datasets with imagery from geographically distributed regions.

However, with the incorporation of DIFL, the out-of-domain performance is greatly enhanced. Analysis criteria includes a comparison of accuracy, sensitivity,

specificity and AUC over both the baseline, as well as the DIFL enhanced algorithms. We conclude that DIFL improves generalizability of Tuberculosis screening while maintaining acceptable accuracy over the source domain imagery when applied across a variety of public datasets.

Keywords: Unsupervised Domain Adaptation, Domain Invariant Feature Learning, TB Screening, Generative Adversarial Networks

REDUCING BIAS IN TUBERCULOSIS SCREENING WITH DEEP DOMAIN ADAPTATION

by

Nishanjan Ravin

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, Baltimore County in partial fulfillment
of the requirements for the degree of
Master of Science
2021

Advisory Committee:
Assistant Professor Dr. David Chapman, Chair/Advisor
Professor Dr. Timothy Finin
Professor Dr. Tim Oates

Acknowledgments

I owe my heartfelt gratitude to all the people who have have supported me in this arduous undertaking, and this thesis would not have been possible without the help and support that I have received over the course of its completion.

First and most importantly, I would like to express my deepest thanks to my advisor, Dr. David Chapman, for granting me the opportunity to work on such a challenging and relevant research problem in the field of Computer Science. His advice and guidance has been invaluable, and his professional expertise and knowledge has led me to successfully overcome the many obstacles that I have faced along the way. It has been an absolute pleasure to work under the direction of such an extraordinary individual, culminating into one of the most enriching experiences in my life thus far.

I would also like to express my gratitude towards Dr. Timothy Finin and Dr. Tim Oates, for agreeing and coordinating to be a part of my thesis defense committee. Their comments and feedback are highly valuable and provide meaningful insights about the subject of my thesis, and I would like to extend my thanks to both of them for taking time off their busy schedule to be participate in my thesis defense.

Next, I would like to thank the people at Rad-AID, Dr. Daniel Mollura, Dr. Farouk Dako, , Ameena Elahi, Alan Schweitzer and Steve Surratt for their encouragement and support on the research work that I have done. Their proficiency in the field of radiology has enabled me to get essential feedback regarding my research,

and their suggestions and discussions have immensely expedited my learning curve, particularly in the field of radiology.

Additionally, I would also like to express my sincere thanks to the people over at the Center of Accelerated Real Time Analytics (CARTA) lab at the University of Maryland, Baltimore County, for the resources that they have provided me throughout the period of my involvement, which allowed me to conduct my research at the fast pace that was expected of me.

Lastly, I would like to thank my family members, as well as my friends, for their priceless support throughout my thesis journey, and for picking me up and motivating me on the days that I did not feel 100%. Their encouragement has enabled me to face the complications and stresses without much difficulty, for which I am always grateful.

It is impossible to remember every single person who have helped me along the way in some way or form, and I apologize to those I've inadvertently left out. Every single help that I have received, either directly or indirectly, have contributed to the completion of my thesis, and is definitely much appreciated by me.

To God be the Glory, The Best is Yet to Be!

Table of Contents

List of Tables

List of Figures

List of Abbreviations and Representations

DIFL	Domain Invariant Feature Learning
TB	Tuberculosis
x_S	Any one source domain image
X_S	The set of all the source domain images
y_S	Any one label of a source domain image
Y_S	The set of all the labels for source domain images
x_T	Any one target domain image
X_T	The set of all the target domain images
y_T	Any one label of a target domain image
Y_T	The set of all the labels for target domain images
d	Any one domain label
D	The set of all the domain labels
$\mathcal{G}$	Generator Network
$\mathcal{C}$	Classifier Network
$\mathcal{D}$	Discriminator Network
α_{C}	Learning rate of label classification step
α_{DI}	Learning rate of domain invariance step

Chapter 1: Introduction

1.1 Domain Adaptation

Generalizability beyond the source domain, i.e. the domain upon which the algorithm is trained on, is an important and difficult challenge for machine learning and deep learning. In medical imaging, it can have a major impact on clinical trustworthiness, as it is unknown whether a deep learning algorithm trained on patients from one population will generalize to another without explicit and extensive out-of-domain testing. All medical imaging datasets contain unique attributes such as patient demographics, imaging procedures, labeling criteria, and the distribution of scanner equipment and settings. It is possible for a machine learning or a deep learning algorithm to overfit these criteria and thereby have difficulty in generalizing to imagery in a new hospital institution or clinic which may have a different distribution of patient population, imaging procedures, equipment, settings, and truth criteria.

One might think that performing cross validation might greatly mitigate the problem of domain shift in medical imagery. However, this is not the case, as both the train and test samples are ultimately drawn from the same dataset with the same patient and imaging distribution. Therefore, it does not provide any indication of

how well an algorithm might perform when applied to a different dataset without access to labeled imagery, even if the target data represents the exact same diagnostic task with a slightly different distribution of imagery.

A canonical example of the aforementioned problem in machine learning is the difficulty of state-of-the-art deep learning networks, trained using the MNIST digits dataset, to accurately predict the USPS zip-code digits dataset without retraining or Domain Adaptation [1]. It is easy for a software developer to erroneously think that a deep learning algorithm trained to predict MNIST digits could be applied whole-cloth without reduced accuracy to handwriting recognition tasks in the wild such as parsing of postal zip codes or handwritten checks, even though the task remains the same.

Fortunately, many recent works have shown that unsupervised Domain Adaptation enables algorithms trained on the MNIST dataset to not only generalize to the USPS dataset, but even to the more difficult Street View House Numbers (SVHN) dataset, which contains images on which digits are typically printed and not handwritten. This is done by first producing generalized features between the different datasets, and then using these generalized features for the classification task at hand. Hence, by using generalized features, the model mitigates the domain shift that is present between the datasets, and can then be used for identical classification tasks on other datasets.

1.2 Tuberculosis Screening

Tuberculosis (TB) is a contagious bacterial infection of the lungs which is widespread globally, thus affecting an estimated 25% of the world's population. 90% of those infected will never have symptoms, but 10% will progress to active TB which frequently causes symptonms such as severe coughing, chest pain, and pulmonary scarring. TB is known to develop more frequently in populations of developing countries, with higher rates of infection in Low and Middle Income Countries (LMIC) than High Income Countries (HIC). In particular, many countries in the African and South American continents have high prevalence of TB infections.

TB can be diagnosed through a few ways, such as conducting a skin test, a blood test, or also by analyzing a chest X-ray image of the patient to look for damage to the lungs. However, the gold standard for diagnosing TB is through conducting a microscopic examination of sputum and culture of bacteria to look for presence of the *Mycobacterium TB*. This golden standard process is a very time-consuming process, which can take up to months at times, and also requires a lot of resources, such as a biosafety level 3 lab. These are not affordable in a large majority of places and countries, and thus diagnosing TB accurately itself proves to a challenge.

It would be desirable to train a deep learning algorithm for TB screening that can generalize to populations from around the world. This would have a huge impact in the medical field, and would particularly find a place in providing much needed assistance for triaging in hospitals across the globe. However, there are 2 major problems that would prevent us from obtaining this ambitious goal.

Firstly, there are not many publicly available TB datasets that can be used to train deep learning algorithms. Deep learning algorithms gain power through looking at large amounts of data, and a lack of accurately annotated TB data would mean that a deep learning algorithm may not perform up to its full potential. Radiologists are able to provide accurate readings of the chest X-ray images to diagnose TB most of the times, which can be used to label TB data if necessary. However, due to the complex nature of the disease, diagnosing TB from just chest X-ray images proves to be a challenging task for even the most experienced radiologists. According to a conducted study [2], it was found that expert radiologists from the top hospitals around the world only have an accuracy of 68.7% in comparison to the gold standard. This is largely due to the fact that the human eye is not sensitive enough to details in the X-ray images to be able to identify TB areas. Apart from difficulty in obtaining gold standard labelled TB data, difficulties in anonymization of patients' data and privacy concerns have led to a lack of public TB datasets that can be used for training deep learning algorithms.

Secondly, the presence of domain shift in TB data would mean that standard deep learning algorithms trained on labelled TB data may not perform well on out-of-domain TB data. For example, a deep learning algorithm trained on a particular TB dataset may only perform well on data from the same TB dataset, and not on data from other TB datasets. In this paper, the impact of domain shift on TB screening algorithms is explored, by utilizing four major public datasets that are available for the identical task of TB screening from chest X-ray images. By training AI models using regular deep learning algorithms on one dataset and cross-

testing them on the other three datasets, we can evaluate the presence of bias (i.e. the inability to generalize to other domains) in these AI models that arise due to domain shift.

Three out of the four public datasets used are regional datasets, with chest X-ray images originating wholly from a singular medical institution in Shenzhen (China), New Delhi (India) and Montgomery County (USA). The last public dataset, referred to as the TBX11K dataset, is a much more extensive dataset, containing a large number of chest X-ray images taken from several medical institutions around the world. As such, one might expect that a deep learning model trained using the TBX11K dataset would be able to perform significantly well on identical TB screening tasks using out-of-domain (in this case, non-TBX11K) data, such as the three regional datasets. However, initial testing has revealed that even state-of-the-art deep learning algorithms, such as the ResNet50 model, when trained only on the TBX11K data, achieves greatly reduced out-of-domain accuracy when tested on the other datasets.

This problem is an ideal candidate for exploring the possibility of applying DIFL methods in the context of medical imaging tasks. As mentioned earlier, DIFL methods have found success in classification tasks such as handwritten digits recognition or object recognition. However, not much research has been done in terms of employing DIFL methods for relatively complex tasks, particularly in the medical field, such as TB screening from chest X-ray images. As such, this paper sets out to explore the viability of improving deep learning algorithms in the medical

field through implementing DIFL methods, by specifically delving into the task of screening TB from chest X-ray images.

1.3 Contributions

The crucial contributions from the research conducted in this thesis are as follows:

1. Assessing and quantifying the presence of bias in standard AI models for screening TB arising due to domain shift
2. Applying and evaluating the effectiveness of the Domain Invariant Feature Learning (DIFL) method for the purpose of TB screening from chest X-ray images across diverse patient populations
3. Discussing the effects of differing disease presentation across TB datasets when employing the DIFL method

1.4 Thesis Statement

Utilizing Domain Invariant Feature Learning (DIFL) can enable a deep Tuberculosis screening model to adapt to changes in disease presentation that can be observed over diverse populations, thus mitigating domain shift.

Chapter 2: Related Work

While there has been interest and developments in the field of domain adaptation recently, it is still an area of machine learning that is yet to be fully explored. Domain adaptation has found its place in a wide variety of applications, ranging from being a part of a broader transfer learning routine [3, 4, 5, 6], explicitly modelling the transformation between two or more domains [7, 8], and even in data augmentation [9, 10]. With a wide variety of flavors to choose from, such as supervised [11, 12, 13], semi-supervised [14, 15, 16], and unsupervised [17, 18, 19], domain adaptation has been increasingly incorporated for an increasing number of tasks, such as handwritten digit classification [20, 21], object recognition in images acquired in different conditions [21, 22], 3D pose estimation [23], and a variety of other tasks.

However, the application of domain adaptation has been relatively limited to slightly simpler tasks, and work done in the field of domain adaptation in the context of medical imaging has been relatively limited thus far. While there has been a slight uptick in utilizing domain adaptation techniques for tasks relating to medical imaging as of recent, none of them have been looked into the field of unsupervised domain adaptation methods. Logically, unsupervised domain adaptation proves to

be the most difficult to achieve in comparison to supervised/semi-supervised, in large part due to the absence of classification labels in the target domain(s). Given the relatively complex nature of machine learning done in the field of medical imaging, i.e. the distribution of data is not as easily learnt as with other simpler tasks such as handwritten digit classification, a large majority of domain adaptation work done in the context of medical imaging has shied away from unsupervised domain adaptation. Some of the closest related work are analysed in the upcoming sections.

[24] looks at how semi-supervised domain adaptation methods can be applied to improve the accuracy of the label classification across different domains for the task of predicting Covid-19 from chest X-ray images. In their proposed semi-supervised Open Set Domain Adversarial (SODA) network, they utilize an adversarial semi-supervised method of training, such that the SODA network will be able to learn features that are adaptable to the target domain. They have used the ChextXray-14 dataset as the source dataset, and the COVID-ChestXray dataset as the target dataset. As this is a semi-supervised approach, they have utilized 40% of the labelled data in the target dataset for training the SODA network. While they have achieved an AUC-ROC score of 0.82 on the target dataset without domain adaptation, they were able to improve upon it by applying domain adaptation, and achieved an AUC-ROC score of 0.90 on the target dataset.

[25], very similarly, looks at the potential of semi-supervised domain adaptation methods to increase the accuracy of label classification on different domains for the task of predicting cardiac abnormalities from chest X-ray images. As with the previous paper, this paper also utilizes a semi-supervised approach, which uses

a limited amount of labelled data from the target dataset as part of training. The source dataset utilized in this paper was the NIH PLCO dataset, and the target dataset was the Indiana University dataset. The authors achieved an accuracy of 0.73 on without domain adaptation, but achieved an improved accuracy of 0.85 with domain adaptation.

[26] has looked at the potential of using domain adaptation methods for the task of lung segmentation from chest X-ray images. They have tested both unsupervised and supervised versions of their domain adaptation algorithm. The Montgomery TB dataset was used as the source, while the JSRT & Pneumoconiosis datasets were used as the target datasets. Without utilizing domain adaptation, they were able to achieve a Dice score of 0.896 and 0.882 on the target datasets respectively. With the supervised version of the domain adaptation, the Dice scores improved to 0.965 and 0.949, while with the unsupervised version of domain adaptation, the Dice score achieved was 0.958 and 0.933.

Next, [27] has also implemented domain adaptation for the task of lung segmentation from chest X-ray images. They have implemented an unsupervised version of domain adaptation, by using the Montgomery TB dataset as the source, and the JSRT dataset as the target. Without domain adaptation, they achieved a Dice score of 0.82 and 0.77 for the right and left lung respectively. However, with unsupervised domain adaptation, the Dice score significantly improved to 0.95 and 0.93.

Lastly, unlike the previous papers we have looked at, [28] looks at applying domain adaptation across data from different modalities, for the task of whole heart

segmentation. They have utilized an unsupervised approach, with the source dataset consisting of MR heart images, and the target dataset consisting of CT heart images. Without domain adaptation, they achieved a Dice score of 0.17, which was greatly improved to a Dice score of 0.73 with domain adaptation.

After looking through the closest relevant work, we are unaware of any research that has looked into the potential of applying unsupervised domain adaptation methods, particularly for the task of TB classification, as we have done in this paper. Most of the work done in the space of unsupervised domain adaptation seems to be related to the task of segmentation, rather than classification. The work done in [24] and [25], as looked at previously, are the closest in terms of research goals and alignment to the work done in this thesis. However, the methods implemented in both of these papers utilize a semi-supervised approach, which assumes the availability of partially labelled data in the target dataset. In reality, such labelled data might not be available in the field of medical imaging (mainly pertaining to chest X-ray images), due to two main factors: difficulty and slowness in obtaining proper ground truths for the chest X-ray images and privacy concerns regarding disclosure of patients' information. As such, un-labelled and un-annotated data is more easily obtainable, which raises the question of whether unsupervised domain adaptation methods would be as effective as semi-supervised domain adaptation methods for the same/similar tasks.

In this paper, we seek to investigate this hypothesis, i.e. how effective would unsupervised domain adaptation methods be in the context of medical imaging

tasks. This is explored by applying unsupervised domain adaptation methods for the particular task of screening for TB from chest X-ray images.

Chapter 3: Methodology

3.1 Overview

In this section, the developed Domain Invariant Feature Learning (DIFL) model will be explained in detail.

The implemented DIFL model draws inspiration from the Generative Adversarial Network (GAN) architecture, as discussed in [29]. The two main components of a basic GAN architecture include the generator network, and the discriminator network. These two networks have opposing goals, with the generator trying to manipulate the input data in a way such that the discriminator fails to successfully categorize the output of the generator into the possible classes, i.e. the generator attempts to "fool" the discriminator, while the discriminator attempts to perform accurate classification. As such, both of these networks have custom loss functions that are adjusted according to the exact function that they serve. A simplified example of the GAN architecture is shown in Figure 3.1.

In the following discussion, an input image is represented by the variable $x \in X$ where X is the set of all input images, and their corresponding classification labels are defined by the variable $y \in Y$, where Y is the set of all classification labels. The set of images and classification labels are also subdivided into source and target

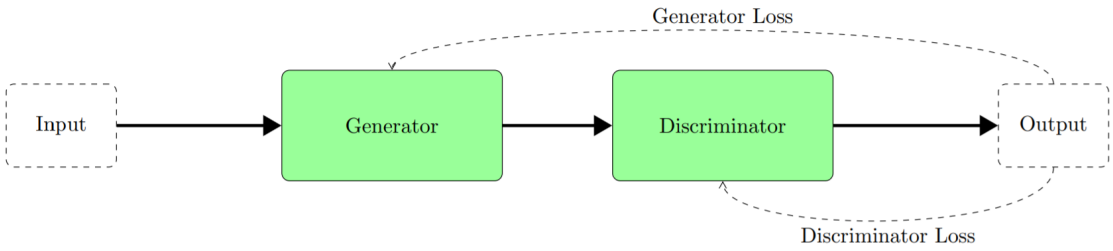

Figure 3.1: Simplified GAN Architecture

domains, which are denoted by the subscript S and the subscript T respectively. As such, we use the variables x_S to denote images from the set of all source domain images X_S, where $x_S \in X_S \subseteq X$, and likewise we use the variable x_T to denote images from the set of all target domain images X_T, where $x_T \in X_T \subseteq X$. Similarly, we define the variables y_S and y_T to denote the classification labels from the set of all source classification labels Y_S and the set of all target classification labels Y_T respectively, where $y_S \in Y_S \subseteq Y$ and $y_T \in Y_T \subseteq Y$. However, it is to be noted that the target classification labels Y_T are unobserved, and hence they are not used for training the DIFL model. Hence, the proposed method in this paper is called **unsupervised DIFL**.

Additionally, to differentiate between the source and the target domains, a domain label is added for each image $x \in X$. The domain labels are denoted by the variable $d \in D$, where D is the set of all domain labels. The variable d_S denotes domain labels taken from the set of all domain labels D_S, such that $d_T \in D_T \subseteq D$, and the variable d_T denotes domain labels taken from the set of all domain labels D_T, such that $d_S \in D_S \subseteq D$. In a similar fashion to classification labels, domain labels can be of any trivial type, as they are only used to differentiate the possible

domains any particular image x could be taken from. Thus, for the purposes of this paper, the source domain labels $d_S \in D_S$ for all source images are set to the value of 0, while the target domain labels $d_T \in D_T$ for all target images are set to the value of 1.

A detailed overview of the DIFL model architecture is shown in Figure 3.2, and explained in the following discussion. The DIFL model, and its training process is explained in the following sections.

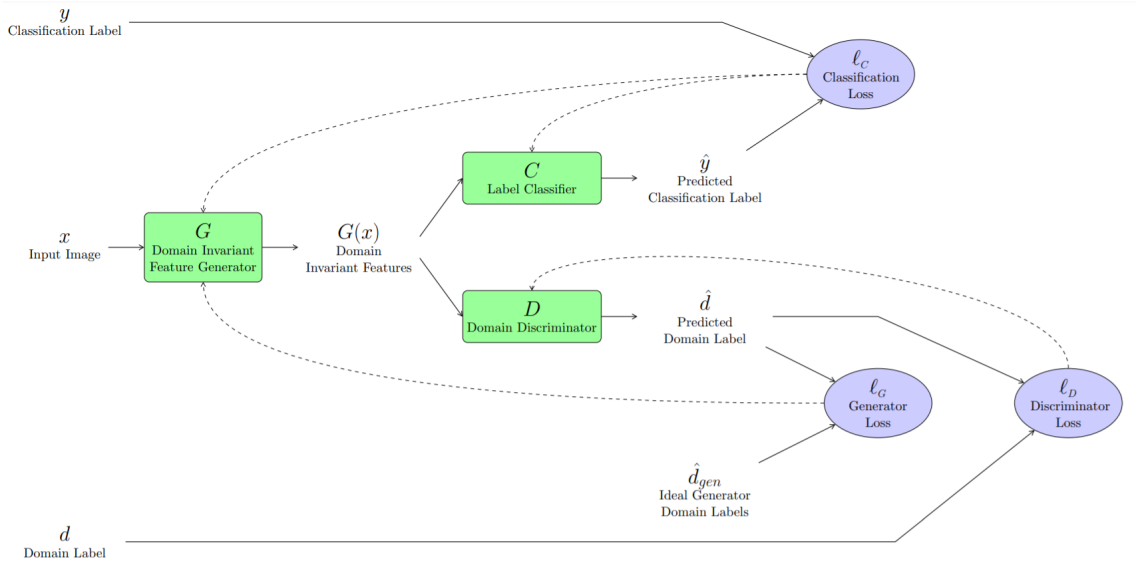

Figure 3.2: Detailed Overview of the DIFL Model

3.2 Training the DIFL Model

The ultimate aim of the DIFL model is to learn generalized feature representations of the images $x \in X$, i.e. from both the source and target domains, while also performing successful label classification on the source images $x_S \in X_S$. This overall

task is accomplished by training 3 separate neural networks simultaneously: a label classifier network, indicated by $\mathcal{C}$, a domain invariant feature generator network, indicated by $\mathcal{G}$, and a domain discriminator network, indicated by $\mathcal{D}$.

In the ideal scenario, the DIFL model would be able to learn and produce perfectly generalizable features of the images $x \in X$, in which case the label classifier network, which has been trained to correctly classify the generalized features of the source images $x_S \in X_S$, can also be utilized to make accurate classifications on the generalized features of the unlabeled target domain images $x_T \in X_T$, as the task is unchanged between the source and target domains. Thus, the DIFL model will be able to perform equivalently well on both source and out-of-domain data, a feat which is not able to be achieved by conventional classification models that are trained using data from a singular domain.

Training the DIFL model would involve training all 3 networks $\mathcal{G}$, $\mathcal{C}$ and $\mathcal{D}$. Each network has its own respective purpose, and hence, is trained through custom methods. Broadly, the training process of the DIFL model can be subdivided into two major steps: a label classification step, and a domain invariance step. The DIFL model is trained by conducting these aforementioned steps simultaneously until convergence of the DIFL model is observed.

3.2.1 Label Classification Step

In the label classification step, image and classification label tuples (x_S, y_S) from the source domain are used to train a part of the DIFL model. The input

images x_S are passed through the domain invariant feature generator network $\mathcal{G}$ to produce $\mathcal{G}(x_S)$, the domain invariant features of the input images x_S. These features are then passed through the label classifier network $\mathcal{C}$, to obtain $\mathcal{C}(\mathcal{G}(x_S))$, which are the predicted classification labels of the images x_S. We define a variable $\hat{y}_S$ to represent the predicted classification labels of the source images, as follows:

$$\hat{y}_S = \mathcal{C}(\mathcal{G}(x_S)) \tag{3.1}$$

The predicted classification labels $\hat{y}_S$ are then compared with the true classification labels y_S, through an appropriate loss function to produce the classification loss, $\ell_{\mathcal{C}}$. Due to the binary nature of the classification labels, the Binary Cross Entropy loss function is utilized to calculate $\ell_{\mathcal{C}}$. This can be represented mathematically using the following formula, where the variable N indicates the total number of images in the batch used for training, and the superscript i is used to denote each individual image:

$$\ell_{\mathcal{C}} = -\frac{1}{N}\sum_{i=1}^{N} y_s^i \times \log\left(\hat{y}_s^i\right) + \left(1 - y_s^i\right) \times \log\left(1 - \hat{y}_s^i\right) \tag{3.2}$$

This loss value $\ell_{\mathcal{C}}$ is used to update the weights of the domain invariant feature generator $\mathcal{G}$ and the label classifier $\mathcal{C}$, by differentiating the loss value with respect to the weights of the respective networks, and then multiplying the resulting gradients with a appropriate learning rate before using the resulting values to update the networks themselves. In the following discussions, this particular learning rate is referred to as the classification learning rate, and represented by the variable α_{C}.

3.2.2 Domain Invariance Step

In the domain invariance step, image and domain label tuples (x, d) from both the source and target domains are used to train a part of the DIFL model. The input images x are passed through the domain invariant feature generator network $\mathcal{G}$ to produce $\mathcal{G}(x)$, the domain invariant features of the input images x. These features are then passed through the domain discriminator network $\mathcal{D}$, to obtain $\mathcal{D}(\mathcal{G}(x))$, which are the predicted classification labels of the images x. We define a variable $\hat{d}$ to represent the predicted domain labels of the images, as follows:

$$\hat{d} = \mathcal{D}(\mathcal{G}(x)), \text{where } x \in X \tag{3.3}$$

As the domain invariant feature generator network $\mathcal{G}$ and domain discriminator network $\mathcal{D}$ are set up in a GAN-like fashion, updating these networks in the domain invariance step is also done in a similar method as with regular GANs. Let us analyze the functions of these two networks in this particular domain invariance step to logically derive the loss functions that should be used to update these two networks.

Regular GANs utilize the minimax loss function to update the generator and discriminator networks, which was first introduced in (Goodfellow et al., 2014), the same paper that proposed the original GAN structure. The minimax loss function, represented by the value function $V(\mathcal{G}, \mathcal{D})$ is as follows:

$$\min_{\mathcal{G}} \max_{\mathcal{D}} V(\mathcal{G}, \mathcal{D}) = \mathbb{E}_x[\log(\mathcal{D}(z_1))] + \mathbb{E}_z[\log(1 - \mathcal{D}(\mathcal{G}(z_2)))] \tag{3.4}$$

While the loss function for the domain invariant feature generator network $\mathcal{G}$ and domain discriminator $\mathcal{D}$ in the domain invariance step would be similar, it need not necessarily be identical to the above minimax loss function. Let us analyze the functions of these networks in closer detail to set up their respective loss functions.

The domain invariant feature generator network $\mathcal{G}$ attempts to produce domain invariant feature representations $\mathcal{G}(x)$ of the source and target images $x \in X$, such that the domain discriminator network $\mathcal{D}$ is unable to correctly identify which domain the images are taken from. The domain discriminator network $\mathcal{D}$ takes $\mathcal{G}(x)$ as input and aims to accurately classify them into their appropriate domains, i.e. correctly predict their domain labels.

Thus, the loss function for the domain discriminator network $\mathcal{D}$ can be set up in a straightforward manner; the predicted domain labels $\hat{d}$ can be compared together with the actual domain labels d through an appropriate loss function to produce a loss value, represented by the variable $\ell_{\mathcal{D}}$. For the purposes of this paper, due to the binary nature of the domain labels (the image can only either belong to the source domain or the target domain), the Binary Cross Entropy loss function is used calculate the loss value $\ell_{\mathcal{D}}$. This can be represented mathematically using the following formula, where the variable N indicates the total number of images in the batch used for training, and the superscript i is used to denote each individual image:

$$\ell_{\mathcal{D}} = -\frac{1}{N}\sum_{i=1}^{N} d_i \times \log\left(\hat{d}_i\right) + \left(1 - d_i\right) \times \log\left(1 - \hat{d}_i\right) \tag{3.5}$$

Setting up the loss function for the domain invariant feature generator network $\mathcal{G}$ requires us to examine its purpose in closer detail. Let us first define the term domain invariant. We can say that the feature representations of a particular image are domain invariant, i.e. the feature representations are highly generalized, when the domain discriminator $\mathcal{D}$ is unable to distinguish which domain the image originates from. Mathematically, this statement can be interpreted in the following manner: the domain discriminator $\mathcal{D}$ assigns an equal probability for the image to be from the source and the target domain.

Thus, the domain invariant feature generator network's goal would be for the domain discriminator network's output, the predicted domain labels $\hat{d}$, to indicate a probability of 0.5 for both the source and target domains. These are termed as the ideal domain labels for the domain invariant feature generator network G, and are represented by the variable $\hat{d}_{\text{gen}}$.

Hence, the loss for the domain invariant feature generator network $\mathcal{G}$, represented by the variable $\ell_{\mathcal{G}}$, can be calculated by comparing the actual predicted domain labels $\hat{d}$ with the ideal domain labels $\hat{d}_{\text{gen}}$ through an appropriate loss function. As with the domain discriminator network, due to the binary nature of the domain prediction subtask, the Binary Cross Entropy loss function is used calculate the loss value $\ell_{\mathcal{G}}$. This can be represented mathematically using the following formula, where the variable N indicates the total number of images in the batch used for training, and the subscript i is used to denote each individual image:

$$
\begin{aligned}
\ell_{\mathcal{G}} &= -\frac{1}{N}\sum_{i=1}^{N} \hat{d}_{\text{gen}} \times \log\left(\hat{d}_i\right) + \hat{d}_{\text{gen}} \times \log\left(1-\hat{d}_i\right) \\
&= -\frac{1}{N}\sum_{i=1}^{N} 0.5 \times \log\left(\hat{d}_i\right) + 0.5 \times \log\left(1-\hat{d}_i\right) \\
&= -\frac{1}{N}\sum_{i=1}^{N} 0.5 \times \left(\log\left(\hat{d}_i\right) + \log\left(1-\hat{d}_i\right)\right)
\end{aligned}
$$

The generator loss value $\ell_{\mathcal{G}}$ is used to update the weights of the domain invariant feature generator network $\ell_{\mathcal{G}}$, while the domain discriminator loss value $l_{\mathcal{D}}$ is utilized to update the weights of the domain discriminator network $\mathcal{D}$. This is done by differentiating the respective loss values with respect to the weights of the corresponding networks, and then multiplying the resulting gradients with an appropriate learning rate before using the resulting values to update the networks themselves. In the following discussions, this particular learning rate is referred to as the domain invariance learning rate, and represented by the variable α_{DI}.

3.2.3 Importance of Hyperparameters

Based on the architecture of the DIFL model, there are many hyperparameters that one would be able to adjust, such as the input image size, batch size, addition of any intermediate layers and many more. However, there are two very crucial parameters that will directly impact performance of the DIFL model; namely, they are the learning rates of the label classification step (α_{C}) and the domain invariance step (α_{DI}) respectively.

It is crucial to find an ideal ratio between α_{C} and α_{DI}, as both of the la-

bel classification and domain invariance steps update the domain invariant feature generator $\mathcal{G}$ directly. The label classification step tries to make sure the domain invariant feature generator $\mathcal{G}$ extracts features that are crucial for label classification, i.e. TB identification, while the domain invariance step tries to ensure that the extracted features are as generalizable as possible between the source and target domains.

These two steps may update the domain invariant feature generator $\mathcal{G}$ in slightly opposing directions, and thus it is crucial to consider the exact ratio between α_{C} and α_{DI}. If the value of α_{C} is set to be too high, while the value of α_{DI} is set to be too low, then we will likely observe a situation where the DIFL model will perform well on the source domain, but perform significantly worse on the target domain. This is due to the fact that the label classification step is given a higher priority than the domain invariance step, and thus the domain invariant feature generator $\mathcal{G}$ will give a higher importance to extracting features relevant to the label classification instead of domain invariance.

Conversely, if the value of α_{C} is set to be too low, while the value of α_{DI} is set to be too high, then we will likely observe a situation where the DIFL model underperforms on the source domain, but achieves a similar performance on both the source and the target domains. This is due to the fact that the domain invariance step is given a higher priority than the label classification step, and thus the domain invariant feature generator $\mathcal{G}$ will give a higher importance to producing generalized features between the source and target domains instead of extracting features relevant to label classification.

Both of these aforementioned scenarios are not ideal, and instead we want a situation where we can maintain a high performance on the source domain, while also maintaining a comparable performance on the target domain. Obtaining the ideal values for these two hyperparameters might involve some initial testing and adjustment, and while the ratio of these two hyperparameters are crucial, the magnitude of the hyperparameters is a factor that we need to keep in mind as well. As such, special attention must be given when choosing the two hyperparameters, α_{C} and α_{DI}.

Chapter 4: Experimental Design

We now detail the experimental design that was undertaken to develop the final domain invariant feature learning (DIFL) model.

4.1 Datasets

The datasets used for the purpose of testing and evaluating the models are listed in this section. All the datasets comprise of chest X-ray scans, which were taken for the primary purpose of detecting TB.

Two public TB datasets were obtained from [30]. The first dataset consists of chest X-ray scans collected at Shenzhen No.3 People's Hospital, Guangdong Medical College, Shenzhen, China, and this dataset consists of 326 healthy cases and 336 TB cases, having a total of 662 X-ray images. The second dataset comprises of chest X-ray scans taken through the cooperation of Department of Health and Human Services, Montgomery County, Maryland, USA, and this dataset has a total of 138 X-ray images, of which 80 are healthy cases and the remaining 58 are TB cases. The third dataset was obtained from [31], which has chest X-ray scans taken from the National Institute of Tuberculosis and Respiratory Diseases, New Delhi, India.

This dataset has a total of 176 chest X-ray images, of which 102 are healthy cases and the remaining 74 are TB cases.

Finally, the fourth dataset to be used was obtained from [2], who had proposed the TBX11K dataset, an extensive dataset comprising of chest X-ray images collected from the top hospitals around the world. However, the TBX11K dataset is divided into 3 categories: healthy, sick but non-TB, and TB cases. For the purposes of this paper, the "sick but non-TB" category has been merged together with the "healthy" category, with the resulting TBX11K dataset containing only 2 classes of images, TB cases and non-TB cases. This is done so as to match with the aforementioned 3 public TB datasets, which also only have 2 binary classifications of the data.

Correspondingly, the TBX11k dataset had 7600 images of non-TB cases, and 800 images of TB cases. Due the high imbalance of data between the two binary classes, data augmentation was done to increase the number of TB images. We applied slight augmentation in the form of rotation with rotation angles limited to 5° so that the appearance of the X-ray images is not severely altered. Through this method, the number of TB images in the TBX11K dataset was increased to 7520 images, which is approximately proportionate to the 7600 non-TB images. Consequently, the final TBX11K dataset which was used has a total of 15120 images.

As a side note, the above method of data augmentation is not ideal, as it may lead to possible data leakages between the train and test splits, wherein augmentated data of the same image may be present in both of the train and test splits of the TBX data. As such, deep learning models will be able to classify such images in

the test splits with ease, as they have already seen the same images in the training split. While this will not drastically impact the performance of the models trained on the TBX dataset, there may be a slight overestimation in performance of the model in the case of testing models on the TBX dataset itself. However, the out-of-domain performance would be unaffected, as data from outside the TBX dataset is not used at all in the training of the models. As the research is mainly focused on out-of-domain performance rather than in-domain performance, this factor does not affect any of the results or discussion presented in the following sections.

Throughout this paper, these aforementioned datasets shall be referred to as the China Dataset, US Dataset, India Dataset and the TBX11K dataset. The datasets that are used for training and testing the models listed in this paper utilize a 80:20 train-test split, both for the source and target domain datasets. Therefore, only 80% of the dataset is used for training, while the other 20% of the data is not seen by the model until the final testing stage.

Samples images of the 4 datasets, from both the TB positive and TB negative classes, are shown in Figure 4.1.

4.2 Evaluation Metrics

Several measures were used to evaluate the performance of the DIFL model and its effectiveness at the task of domain adaptation. These measures are detailed in the following discussion.

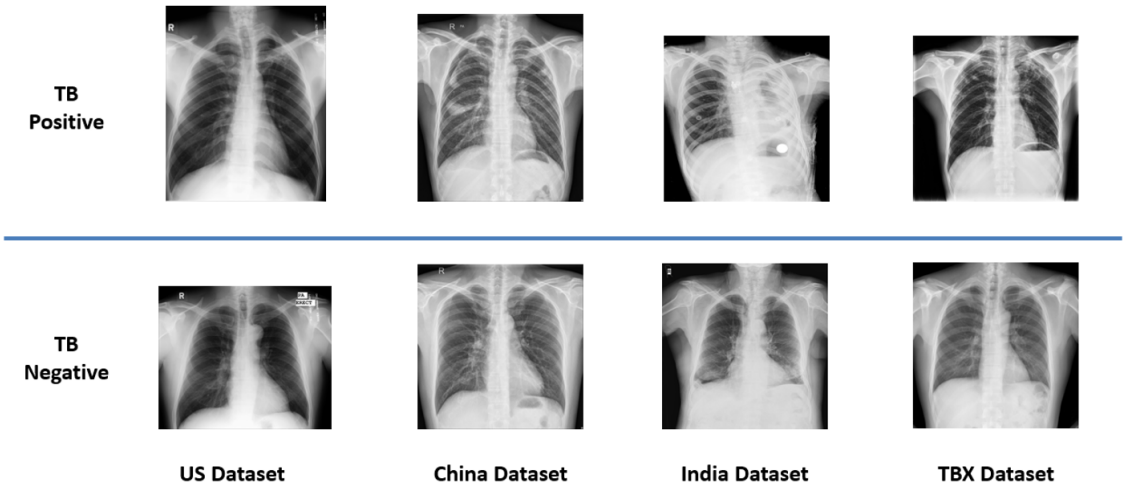

Figure 4.1: Sample of Chest X-ray images from the four respective datasets

4.2.1 Accuracy

Accuracy of label classification is one of the most crucial measures of evaluating the DIFL model, as it can give us a direct idea of how well it can perform in its primary task of TB prediction. Calculating the accuracy of the DIFL model is straightforward. Input images and their corresponding classification labels (x, y) are sampled from any domain, i.e. either source or target domain. The input images are passed through the domain invariant feature generator network $\mathcal{G}$, and to produce the domain invariant features, $\mathcal{G}(x)$. These features are then passed through the label classifier network $\mathcal{C}$, to produce $\mathcal{C}(\mathcal{G}(x))$, the predicted classification label, which can also be represented as $\hat{y}$. This predicted label is compared with the true classification label y.

The above process is repeated for all data in the domain, and a confusion matrix is built, consisting of the number of true positives (TP), false positives (FP),

true negatives (TN) and false negatives (FN). The accuracy of the DIFL model for this particular domain can then be calculated using to the following formula:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4.1}$$

The resulting accuracy value would be a great indicator of how well the DIFL model is able to correctly classify the input data. However, considering a single performance metric would not be a good enough basis to fully evaluate a model's performance. Hence, additional metrics are also used to supplement our analysis.

4.2.2 Sensitivity (True Positive Rate) and Specificity (True Negative Rate)

Another metric which can be utilized to analyze the DIFL model is sensitivity. Evaluating the sensitivity can be done in a similar fashion as with accuracy, by first building the confusion matrix. The sensitivity value is then calculated using the following formula:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.2}$$

The sensitivity value provides a good measure of how well the model can correctly classify the TB positive instance. Essentially, it represents the proportion of TB positive cases that were correctly predicted as TB positive by the model as well, and hence this value is also known as the true positive rate.

Apart from sensitivity, the specificity is also a possible metric that can be

used to evaluate the DIFL model. Using the confusion matrix, the specificity value is calculated using the following formula:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4.3}$$

The specificity value is similar to the sensitivity value, but instead indicates how well the model can correctly classify the TB negative instances, i.e. it represents the proportion of TB negative cases that were correctly predicted as TB negative by the model. Thus, this value is also known as the true negative rate.

4.2.3 ROC - AUC Measure

The Receiver Operating Characteristic (ROC) curve is a graphical plot that visualizes the classification ability of a binary classifier model, by varying its discrimination threshold and evaluating its performance on the data. This is done by plotting a graph of its True Positive Rate (Sensitivity) against its False Positive Rate (1-Specificity) as its discrimination threshold is varied from 0 to 1. While the ROC curve enables one to visualize a model's performance, the graph alone does not however provide a concrete method to objectively compare the performance of different models.

To supplement the analysis of a model's ROC curve, an additional metric, known as the Area Under Curve (AUC) value, can be calculated and used to judge the performance of the binary classifier model. The AUC value can be determined trivially by taking the area under the ROC curve (this can be obtained by integrating

with respect to the ROC curve), and provides a simple but effective way of directly comparing the performances of different models.

As the True Positive Rate and False Positive Rate ranges from a minimum value 0 to a maximum value of 1, the maximum AUC value possible would be $1 \times 1 = 1$, which is only obtainable by a model that is able to perform perfect binary classification at all possible discrimination thresholds. A random binary classifier is expected to achieve an AUC value of 0.5, as it can perform correct predictions about approximately only half of the time. As the AUC value tends closer to 1, we can infer that the model is able to better classify the instances of the data, and hence is indicative of better model performance.

Chapter 5: Results and Discussion

5.1 Overview

In this section, the obtained results from the aforementioned experiments will be detailed and discussed.

In order to assess and quantify the domain shift in TB datasets when utilizing standard deep learning algorithms, two non-DIFL models were implemented. These non-DIFL models were trained on a single dataset of the four TB datasets, and then tested on all four datasets to evaluate their performance on in-domain testing and out-of-domain testing.

Firstly, a simple non-DIFL model was trained on each of the four datasets. The simple non-DIFL model's architecture constists of a single Convolution Neural Network, which has a couple of convolutional layers followed by some dense layers. The results obtained from evaluating the performance of this simple non-DIFL model on all four of the datasets are presented in Figures 5.1-5.4, and Tables 5.1-5.4.

Next, a more advanced non-DIFL model was trained on each of the four datasets. The advanced non-DIFL model's architecture consists of a ResNet50 network at the start, a VGG19 network in the middle, and a few dense layers at the end. The results obtained from evaluating the performance of this advanced non-

DIFL model on all four of the datasets are presented in Figure 5.5-5.8, and Tables 5.5-5.8.

Lastly, the DIFL models were implemented, which are described in further detail in the upcoming sections.

5.2 Simple Non-DIFL Model Results

5.2.1 Simple Non-DIFL Model: Trained on China Dataset

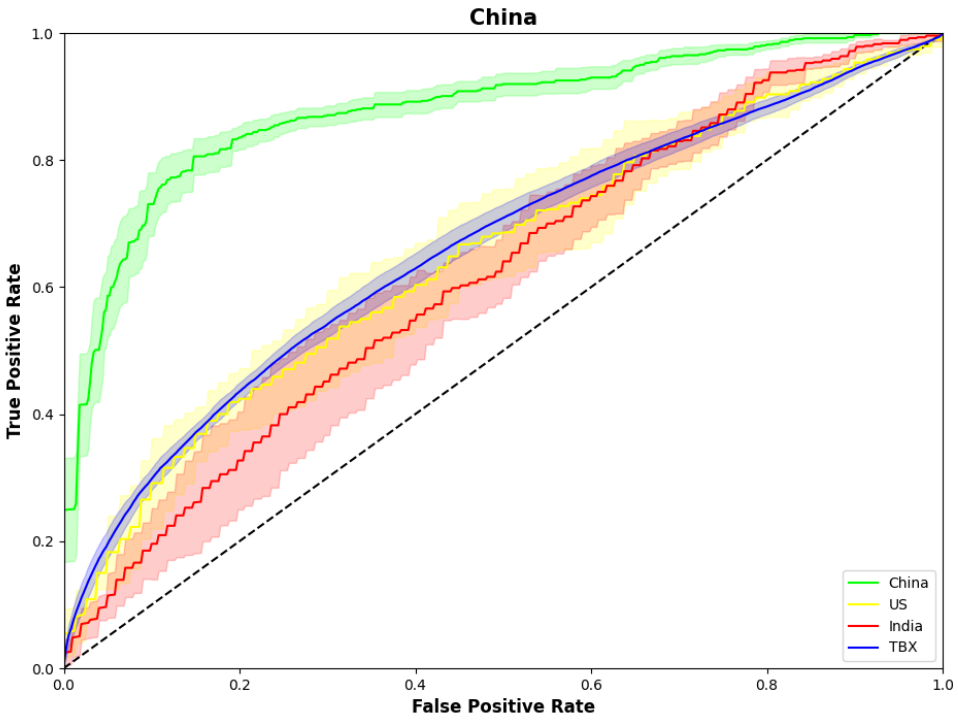

Figure 5.1: ROC Curve: Simple Non-DIFL Model trained on China Dataset

Dataset	**Accuracy**	**AUC**
China	0.84 ± 0.01	0.88 ± 0.01
India	0.55 ± 0.02	0.62 ± 0.04
US	0.57 ± 0.03	0.65 ± 0.04
TBX	0.59 ± 0.01	0.66 ± 0.01

Table 5.1: Accuracy and AUC: Simple Non-DIFL Model trained on China Dataset

5.2.2 Simple Non-DIFL Model: Trained on India Dataset

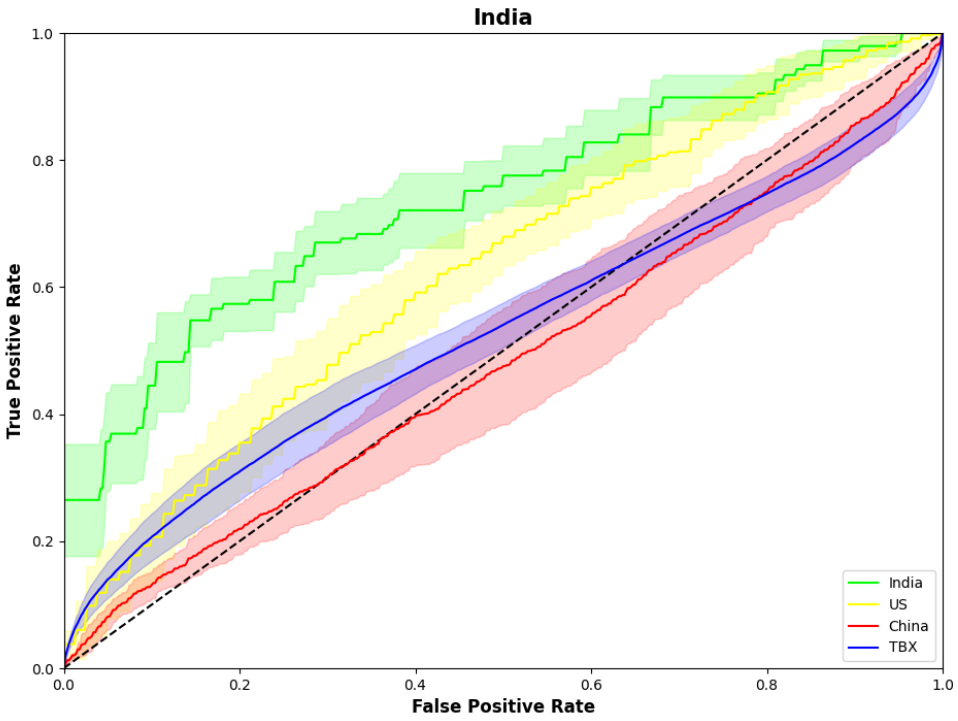

Figure 5.2: ROC Curve: Simple Non-DIFL Model trained on India Dataset

Dataset	**Accuracy**	**AUC**
China	0.49 ± 0.03	0.48 ± 0.05
India	0.77 ± 0.02	0.74 ± 0.03
US	0.60 ± 0.03	0.63 ± 0.04
TBX	0.54 ± 0.02	0.53 ± 0.03

Table 5.2: Accuracy and AUC: Simple Non-DIFL Model trained on India Dataset

5.2.3 Simple Non-DIFL Model: Trained on US Dataset

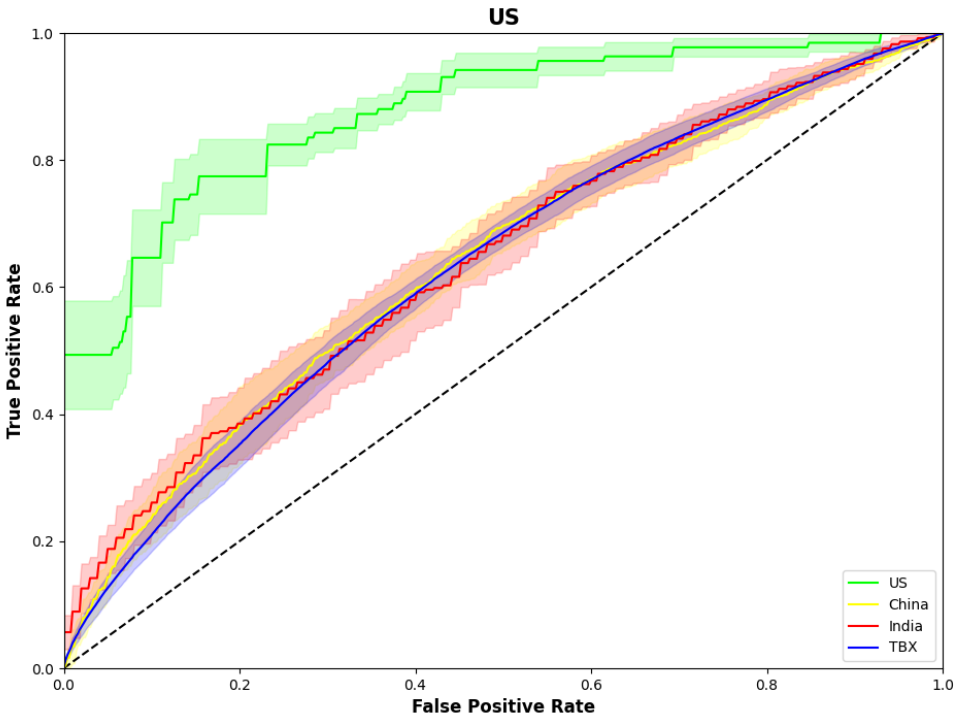

Figure 5.3: ROC Curve: Simple Non-DIFL Model trained on US Dataset

Dataset	**Accuracy**	**AUC**
China	0.53 ± 0.02	0.64 ± 0.03
India	0.46 ± 0.03	0.64 ± 0.03
US	0.86 ± 0.04	0.88 ± 0.04
TBX	0.53 ± 0.02	0.63 ± 0.02

Table 5.3: Accuracy and AUC: Simple Non-DIFL Model trained on US Dataset

5.2.4 Simple Non-DIFL Model: Trained on TBX Dataset

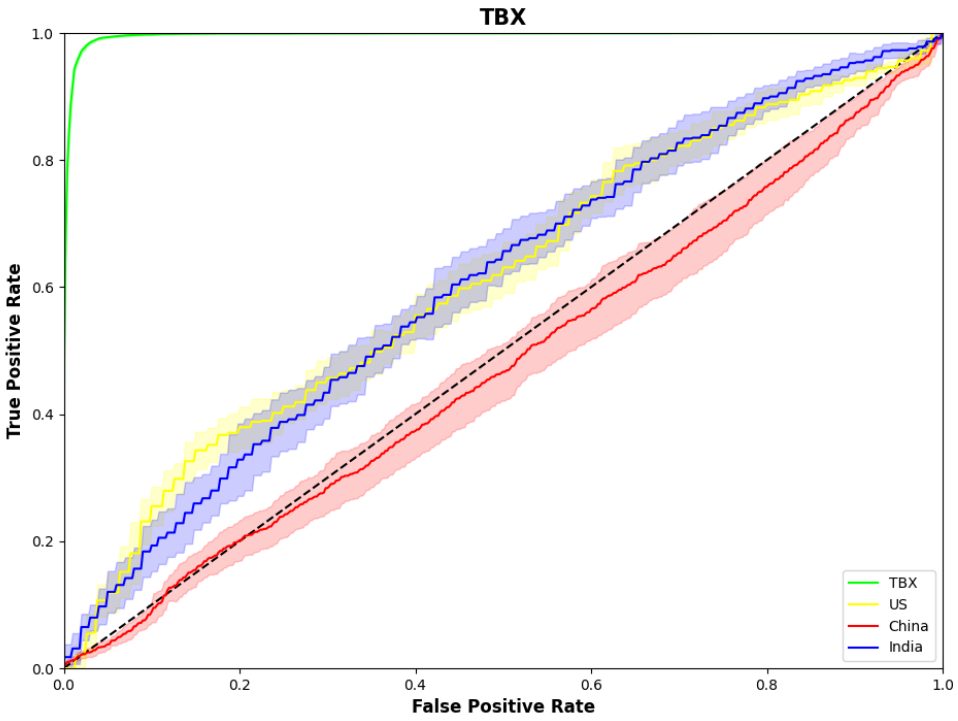

Figure 5.4: ROC Curve: Simple Non-DIFL Model trained on TBX Dataset

Dataset	**Accuracy**	**AUC**
China	0.48 ± 0.02	0.48 ± 0.03
India	0.58 ± 0.01	0.61 ± 0.02
US	0.57 ± 0.02	0.61 ± 0.01
TBX	0.98 ± 0.01	0.99 ± 0.01

Table 5.4: Accuracy and AUC: Simple Non-DIFL Model trained on TBX Dataset

5.3 Advanced Non-DIFL Model Results

5.3.1 Advanced Non-DIFL Model: Trained on China Dataset

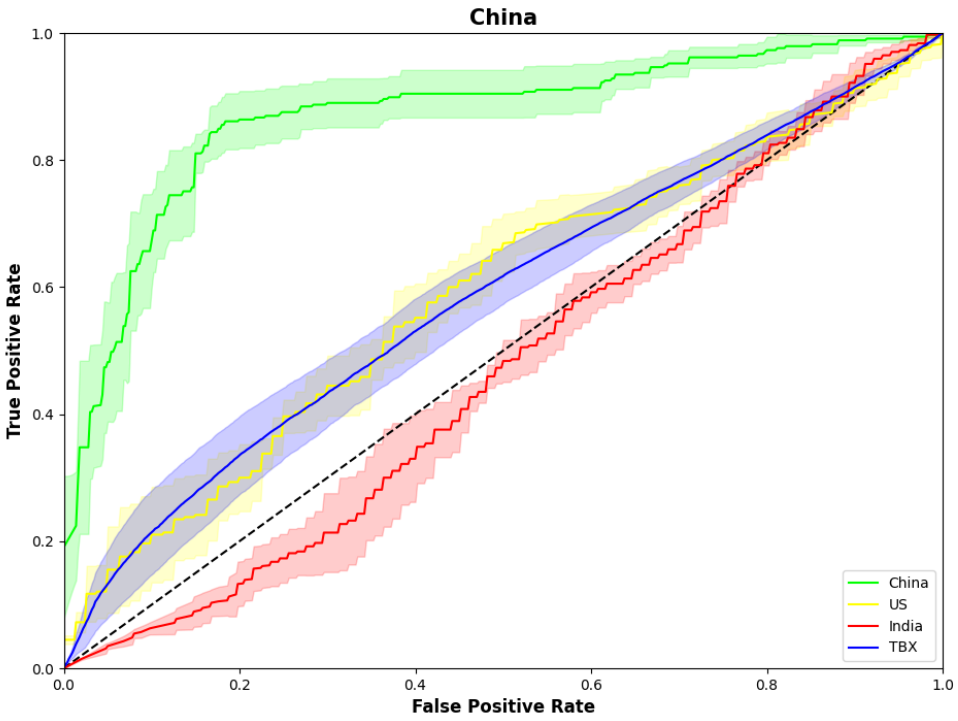

Figure 5.5: ROC Curve: Advanced Non-DIFL Model trained on China Dataset

Dataset	**Accuracy**	**AUC**
China	0.84 ± 0.02	0.87 ± 0.02
India	0.46 ± 0.01	0.47 ± 0.02
US	0.59 ± 0.01	0.59 ± 0.03
TBX	0.56 ± 0.02	0.59 ± 0.04

Table 5.5: Accuracy and AUC: Advanced Non-DIFL Model trained on China

5.3.2 Advanced Non-DIFL Model: Trained on India Dataset

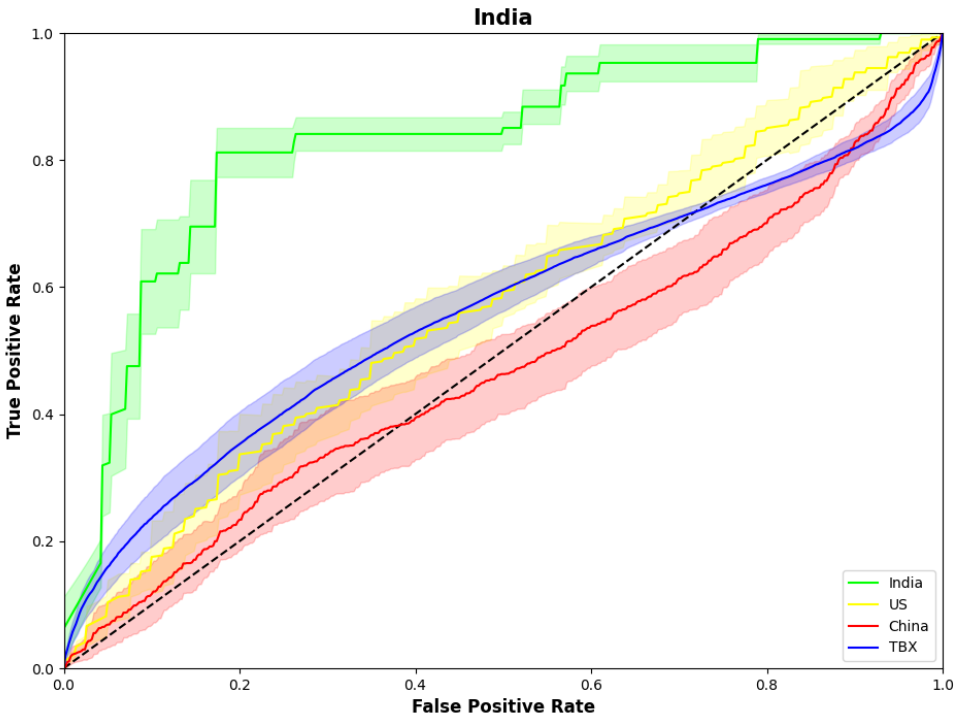

Figure 5.6: ROC Curve: Advanced Non-DIFL Model trained on India Dataset

Dataset	**Accuracy**	**AUC**
China	0.52 ± 0.03	0.47 ± 0.04
India	0.83 ± 0.02	0.83 ± 0.02
US	0.53 ± 0.05	0.56 ± 0.02
TBX	0.57 ± 0.02	0.57 ± 0.04

Table 5.6: Accuracy and AUC: Advanced Non-DIFL Model trained on India Dataset

5.3.3 Advanced Non-DIFL Model: Trained on US Dataset

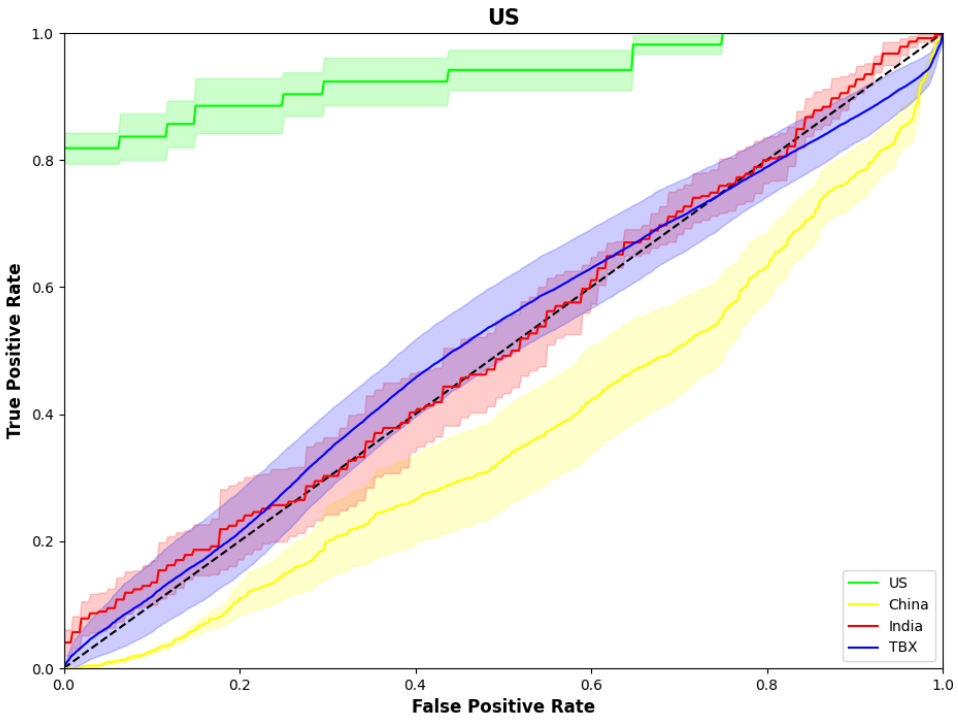

Figure 5.7: ROC Curve: Advanced Non-DIFL Model trained on US Dataset

Dataset	**Accuracy**	**AUC**
China	0.46 ± 0.01	0.37 ± 0.05
India	0.50 ± 0.03	0.51 ± 0.03
US	0.93 ± 0.01	0.93 ± 0.02
TBX	0.48 ± 0.02	0.52 ± 0.05

Table 5.7: Accuracy and AUC: Advanced Non-DIFL Model trained on US Dataset

5.3.4 Advanced Non-DIFL Model: Trained on TBX Dataset

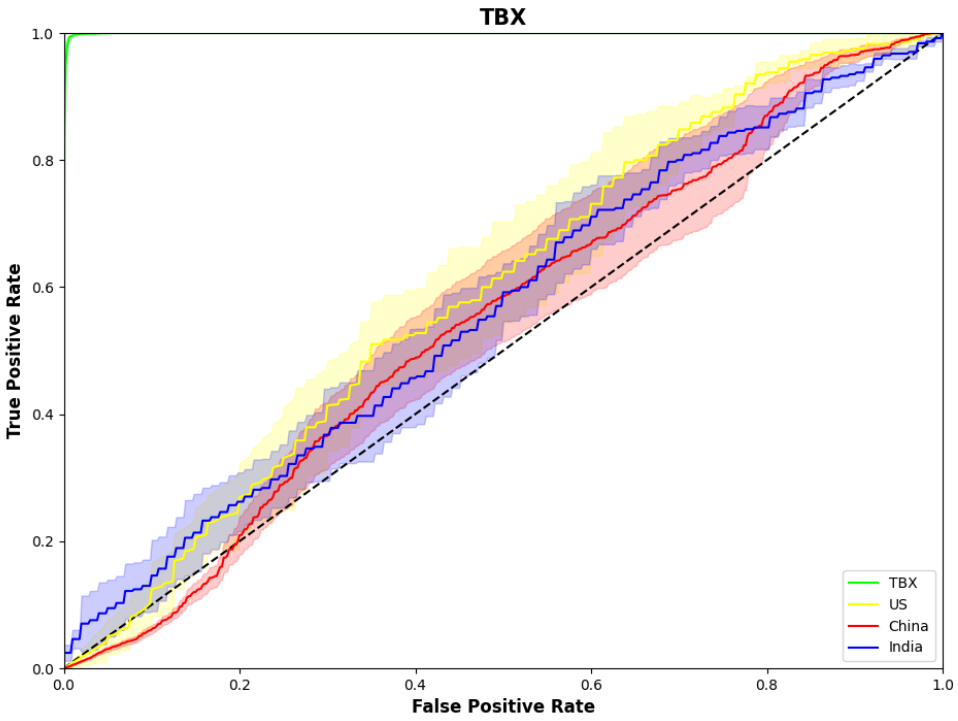

Figure 5.8: ROC Curve: Advanced Non-DIFL Model trained on TBX Dataset

Dataset	**Accuracy**	**AUC**
China	0.56 ± 0.02	0.55 ± 0.04
India	0.54 ± 0.04	0.56 ± 0.04
US	0.48 ± 0.04	0.59 ± 0.05
TBX	0.99 ± 0.01	0.99 ± 0.01

Table 5.8: Accuracy and AUC: Advanced Non-DIFL Model trained on TBX Dataset

5.4 Non-DIFL Model Results - Discussion

From the results obtained through both the simple and the advanced non-DIFL models, it is evident that standard deep learning models, regardless of how advanced they are, are unable to generalize to other domains.

In all the cases seen above, the non-DIFL model achieves good performance on the dataset it was trained on, as evident from the ROC curve and the significantly higher accuracy and AUC scores, with values as high as around 0.8. However, when tested on the other datasets, the performance drops close to random guessing, with accuracy and AUC scores in the region of 0.5-0.6, and the ROC curve observed to be close to the diagonal. This confirms the presence of domain shift and its effect of preventing conventional models to fail when tested on datasets other than the one they have been trained on.

One interesting observation is that even standard models trained on the TBX dataset, which is a significantly larger dataset with images from hospitals all over the world, are unable to generalize to the regional datasets. (While the in-domain performance of the non-DIFL models trained on the TBX dataset might be slightly overestimated, as noted in Section 4.1, the out-of-domain performance would not be affected, which is the key point in this discussion.) As such, we can infer that population diversity is not the only factor that contributes to domain shift. If that was the case, it is expected that a standard model trained on the TBX dataset would perform well on the three other regional datasets.

Thus, the above set of results highlight the need for domain adaptation meth-

ods, particularly in the case where out-of-domain performance and capability of generalization across differing datasets is necessary.

5.5 DIFL Model Results

In this section, we look at the results obtained from implementing the DIFL approach.

As part of the experimentation, the three regional datasets were used, with each of the China, India and US datasets being utilized as the source dataset, while one of the remaining datasets was used as the target dataset. There are six such possible combinations of source and target datasets by using these three regional datasets. For the sake of simplifying labels, the $X \rightarrow Y$ nomenclature is used to define the source and target datasets, where X is the source dataset, and Y is the target dataset.

In each trial per combination, three different types of models are trained and evaluated accordingly: the first model is a non-DIFL model, which is trained on the source dataset, and then tested on the target dataset. This model, as expected, does not achieve good performance measures, due to the presence of domain shift. As such, the performance scores achieved by this model are used as a minimum baseline, and hence this model is termed as the *lower baseline model*.

The second model is a non-DIFL model, wherein the model is directly trained on the target dataset, and consequently tested on the target dataset. As it is being directly trained on the data upon which it is also tested, this model is expected to

perform well. The performance scores from this model provide an "upper bound" to which the results from the experimental DIFL model can be compared against, and thus this model is termed as the *upper baseline model*.

The final model is the DIFL model, which is trained on the source dataset, and tested on the target dataset utilizing the DIFL algorithm. The performance of this model can be evaluated by comparing it against the previously mentioned lower and upper baseline models.

5.5.1 Source Domain: China, Target Domain: India

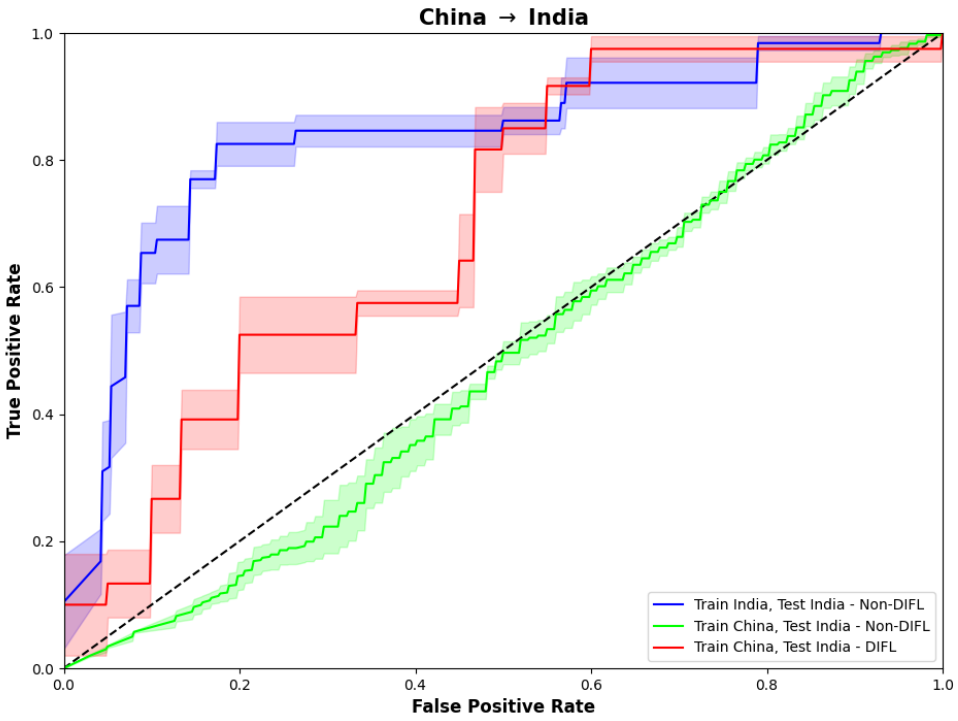

Figure 5.9: ROC Curve: Source Domain - China, Target Domain - India

Type of Model	**Accuracy**	**AUC**
Lower Baseline Model	0.47 ± 0.01	0.48 ± 0.01
DIFL Model	0.68 ± 0.02	0.70 ± 0.01
Upper Baseline Model	0.84 ± 0.01	0.83 ± 0.02

Table 5.9: Accuracy and AUC: Source Domain - China, Target Domain - India

5.5.2 Source Domain: China, Target Domain: US

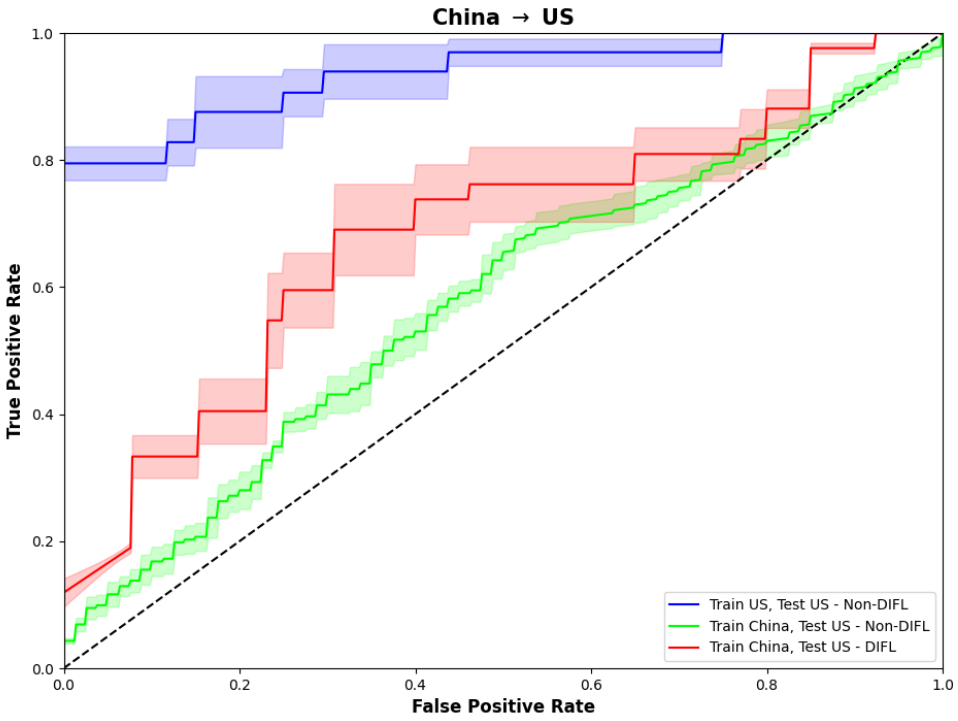

Figure 5.10: ROC Curve: Source Domain - China, Target Domain - US

Type of Model	**Accuracy**	**AUC**
Lower Baseline Model	0.59 ± 0.01	0.57 ± 0.02
DIFL Model	0.71 ± 0.01	0.68 ± 0.04
Upper Baseline Model	0.93 ± 0.02	0.93 ± 0.02

Table 5.10: Accuracy and AUC: DIFL Model Source Domain - China, Target Domain - US

5.5.3 Source Domain: India, Target Domain: China

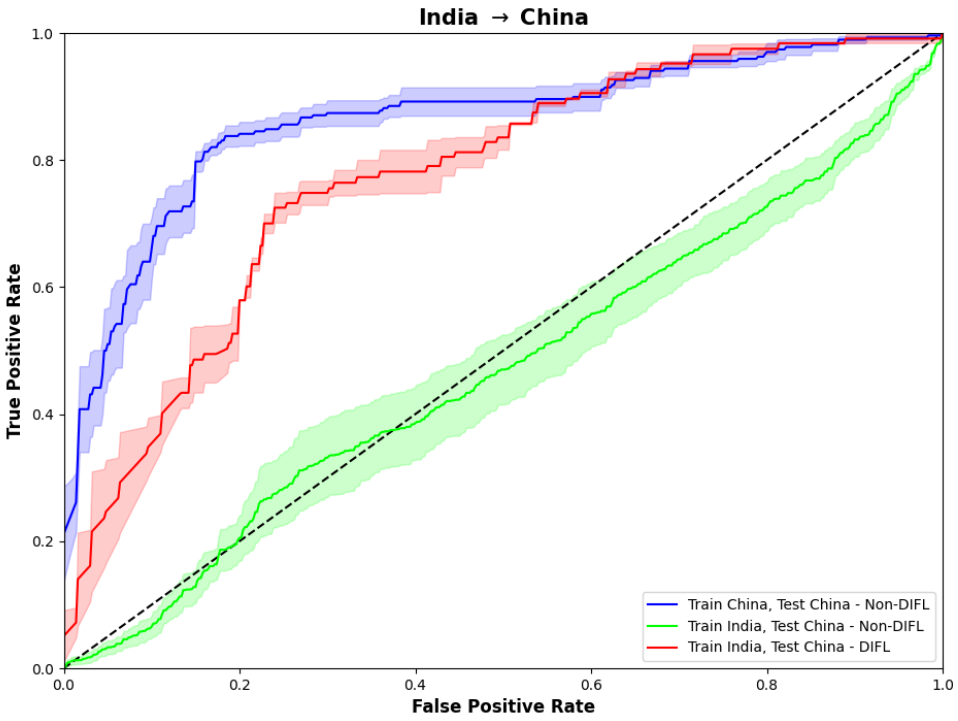

Figure 5.11: ROC Curve: Source Domain - India, Target Domain - China

Type of Model	**Accuracy**	**AUC**
Lower Baseline Model	0.50 ± 0.02	0.47 ± 0.04
DIFL Model	0.73 ± 0.02	0.77 ± 0.01
Upper Baseline Model	0.83 ± 0.01	0.86 ± 0.01

Table 5.11: Accuracy and AUC: Source Domain - India, Target Domain - China

5.5.4 Source Domain: India, Target Domain: US

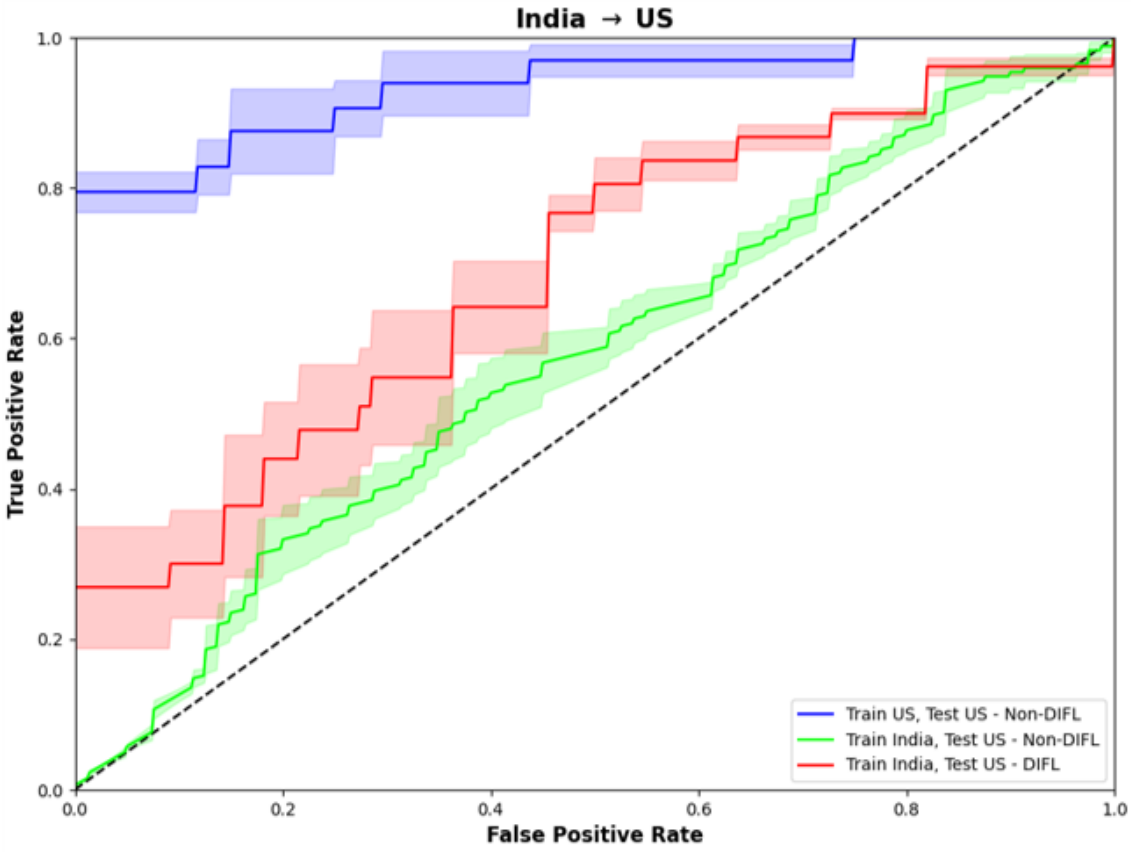

Figure 5.12: ROC Curve: Source Domain - India, Target Domain - US

Type of Model	**Accuracy**	**AUC**
Lower Baseline Model	0.49 ± 0.03	0.57 ± 0.02
DIFL Model	0.63 ± 0.04	0.68 ± 0.04
Upper Baseline Model	0.93 ± 0.02	0.93 ± 0.02

Table 5.12: Accuracy and AUC: Source Domain - India, Target Domain - US

5.5.5 Source Domain: US, Target Domain: China

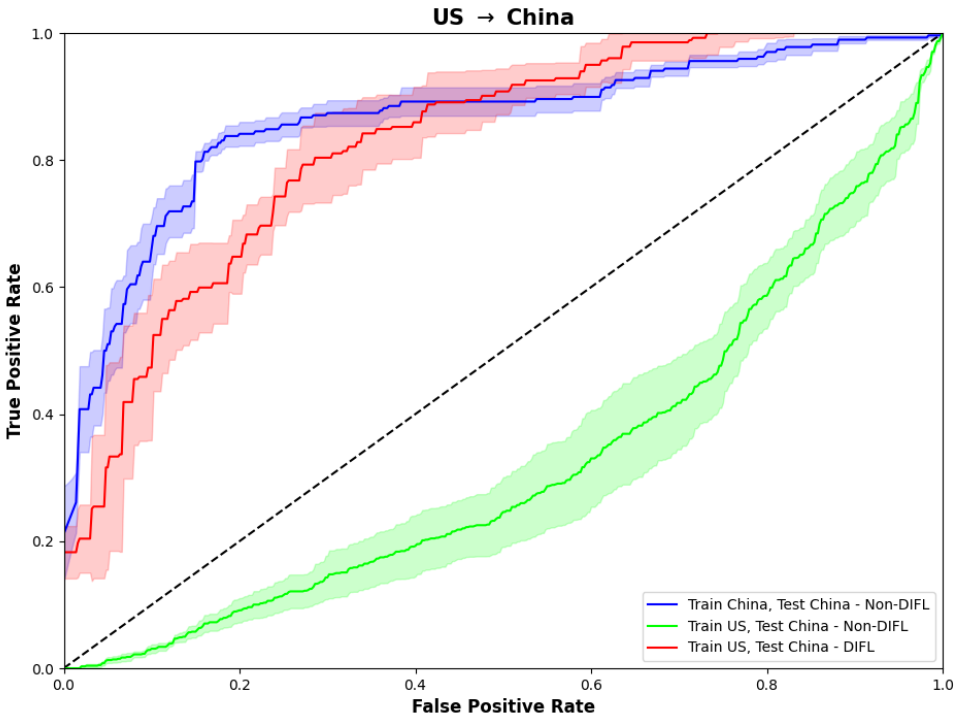

Figure 5.13: ROC Curve: Source Domain - US, Target Domain - China

Type of Model	**Accuracy**	**AUC**
Lower Baseline Model	0.45 ± 0.01	0.33 ± 0.04
DIFL Model	0.80 ± 0.01	0.84 ± 0.04
Upper Baseline Model	0.82 ± 0.01	0.86 ± 0.01

Table 5.13: Accuracy and AUC: Source Domain - US, Target Domain - China

5.5.6 Source Domain: US, Target Domain: India

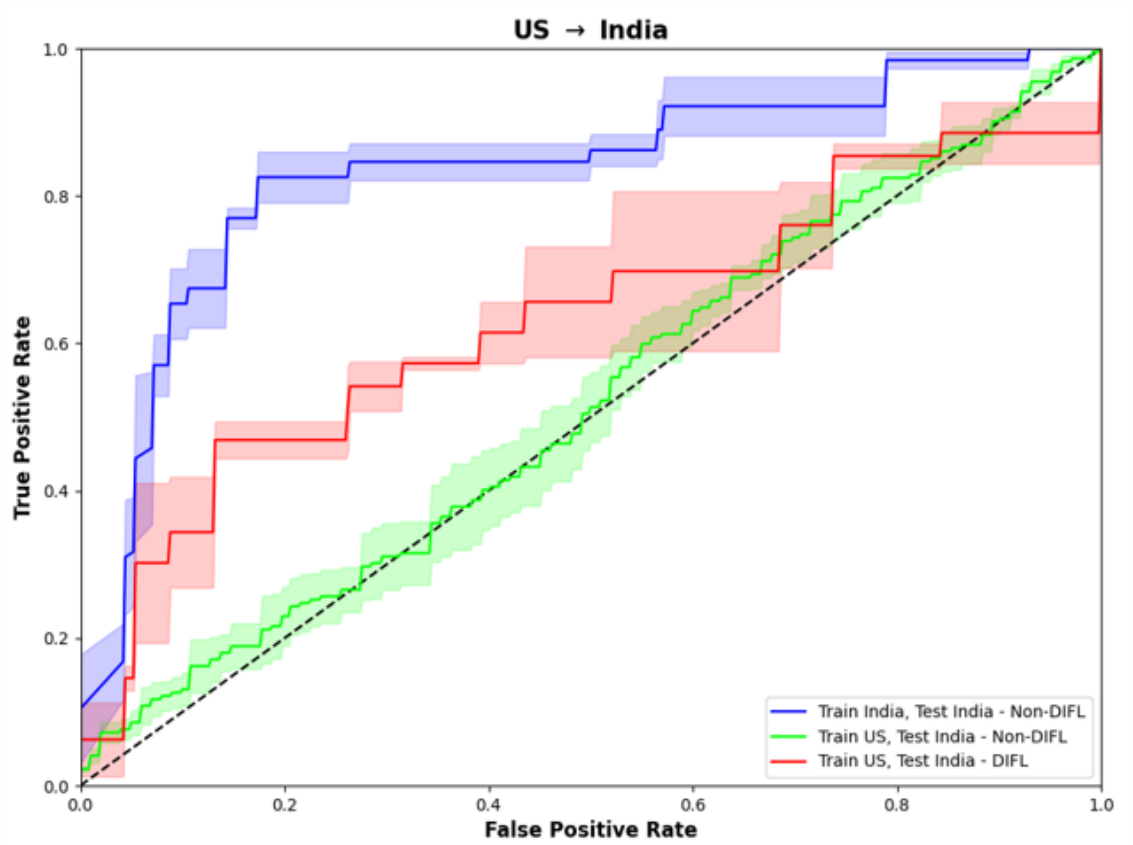

Figure 5.14: ROC Curve: Source Domain - US, Target Domain - India

Type of Model	**Accuracy**	**AUC**
Lower Baseline Model	0.52 ± 0.02	0.52 ± 0.03
DIFL Model	0.63 ± 0.02	0.63 ± 0.02
Upper Baseline Model	0.83 ± 0.01	0.83 ± 0.02

Table 5.14: Accuracy and AUC: Source Domain - US, Target Domain - India

5.6 DIFL Model Results - Discussion

In the following sections, the DIFL model results presented above are explained in detail.

5.6.1 Accuracy Scores

The accuracy scores of all three of these models, for each of the six possible combinations amongst the regional datasets, are detailed in Tables 5.9-5.14.

It is observed that in all the six combinations, the non-DIFL lower baseline model is only able to achieve accuracy scores of around 0.5, signifying that these models are not performing any better than random guessing. The upper baseline model, as expected, achieves high accuracy scores in the region of 0.8-0.9.

Looking at the DIFL models, it is observed that they achieve significantly higher accuracy scores than their respective lower baseline models, by approximately 0.2. While it is not able to achieve accuracy scores as high as the upper baseline models, it does come close in one combination, particularly the US $\rightarrow$ China case, wherein the DIFL model achieves an accuracy score of 0.80 (which is a great increase from the lower baseline model's accuracy of 0.45), while the upper baseline model achieves an accuracy score of 0.82. The case of India $\rightarrow$ China also achieves a relatively good DIFL model performance, achieving an accuracy score of 0.73, which is a significant increase from the lower baseline model's accuracy score of 0.50, and falling just short of the upper baseline model's accuracy score of 0.83.

Hence, while the DIFL model is not able to outperform the upper baseline

model, it can be said that in some cases, it has the capability to achieve a similar performance as that of the upper baseline model in certain conditions.

5.6.2 ROC Curves

ROC curves can further aid in visualizing and comparing the performance amongst the three types of models stated above. These ROC curves are presented in Figures 5.9-5.14.

Each ROC graph corresponds to a particular combination of the regional datasets. In the graphs, a consistent color scheme is utilized for easier comparison and analysis. The lower baseline model is represented by the green curve, the upper baseline model is represented by the blue curve, and the DIFL model is represented by the red curve. Additionally, a region of one standard deviation is also shaded in a lighter color to represent the margin of deviation of each curve.

It is observed that in all six cases, the non-DIFL lower baseline model has a ROC curve that is rather similar to the diagonal, which is once again indicative of the fact that the lower baseline model is only able to achieve a performance that is similar to random guessing. The upper baseline model is able to achieve good results, as indicated by the ROC curve's shape which is highly directed towards the top left corner.

Looking at the DIFL model's ROC curve, it is observed that the DIFL model is able to significantly outperform the non-DIFL lower baseline model, as noted by the significant shift in the ROC curve towards the upper left corner. When

comparing the DIFL model's ROC curve with the upper baseline model's ROC curve, the latter seems to perform better in most cases. However, in one particular combination, namely the US → China case, the DIFL model's performance seems to be on par with the upper baseline model's performance. The case of India → China also has comparable performances between the DIFL model and the upper baseline model. This observation corroborates the evidence that the DIFL model has the potential to perform as well as the model that was directly trained on the target dataset.

Thus, analyzing the ROC curves brings more evidence to the fact that the DIFL model always significantly outperforms the lower baseline model, and in certain conditions, can perform as well as the upper baseline model as well.

5.6.3 AUC Scores

AUC scores are used to quantify the performance of the models from the ROC curves. The AUC scores of all three models, for each of the six possible combinations amongst the regional datasets, are detailed in Tables 5.9-5.14.

It is observed that in all the six combinations, the non-DIFL lower baseline model is only able to achieve AUC scores of around 0.5, signifying that these models are not performing any better than random guessing. The upper baseline model, as expected, achieves high AUC scores in the region of 0.8-0.9.

Looking at the DIFL models, it is observed that they achieve significantly higher AUC scores than their respective lower baseline models, by approximately

0.1-0.2. While it is not able to achieve AUC scores higher than the upper baseline models, it does come close in one combination, particularly the US → China case, wherein the DIFL model achieves an AUC score of 0.84 (great increase from the lower baseline model's AUC score of 0.33), while the upper baseline model achieves an AUC score of 0.86. Also, the India → China case also achieves relatively good DIFL model performance, achieving a AUC score of 0.77 (significant increase from the lower baseline model's AUC score of 0.47), which is just slightly less than the upper baseline model's AUC score of 0.86.

Thus, evaluating the AUC scores brings us to a similar conclusion as when evaluating on the previous performance metrics. While the DIFL model is not able to outperform the upper baseline model, it can be said that in some cases, it has the capability to achieve a similar performance as that of the upper baseline model in certain conditions.

5.6.4 Discussion of Disease Presentation

One interesting observation from the conducted experiments was the case of using the US dataset as the source dataset, and the Chinese dataset as the target dataset. This particular combination saw the highest improvement in performance from the lower baseline model, and the closest performance in comparison to the upper baseline model. Another combination that produced relatively good results was the case where the India dataset was used as the source dataset, and the China dataset was used as the target dataset. As such, we notice that the DIFL algorithm

is able to generalize particularly well to the China dataset. This can be attributed to differences in disease presentation among the three regional datasets.

There are a variety of factors that can affect disease presentation in patients, particularly in the case of screening for TB from chest X-ray images. These factors include, but are not limited to: X-ray exposure, X-ray quality, position of the patient, and other medical equipment configurations.

Additionally, it is known that TB is a progressive disease, and as such, there may be differing levels of manifestations in patients affected with TB. Some TB patients may have a more serious manifestation of TB, while others may have a less severe manifestation of TB. As such, the severity of the disease itself, can cause differing disease presentation in patient populations, as it will directly affect the extent to which the lung is damaged as a result of being affected by TB.

The varying disease presentation in each of the three regional datasets is the main reason as to why we observe relatively better results when employing the DIFL method and using the China dataset as the target dataset. Upon further inspection of the chest X-ray images by certified radiologists, it was observed that some of the images in the China dataset do have different disease presentation of TB, when compared with the other datasets. While all of the three regional datasets themselves had varying disease presentations, the China dataset had disease presentation that was largely different from the images in the US or India datasets.

Thus, to achieve ideal performance of the DIFL model, the disease presentation should be considered when choosing the source and target datasets. Certain source datasets allow for better extracting of generalized features, and thus can achieve

better generalization across target datasets. As evidenced by the results we saw before, by analyzing disease presentation and choosing the right combination of the source and target datasets, the DIFL model would have the potential to produce ideal results that are comparable to training directly on labelled data from the target dataset.

As such, it is crucial to consider the nature of the datasets, particularly in the task of screening for TB from chest X-ray images, when employing the DIFL approach. Due to the complex nature of TB, the disease presentation factor may affect the effectiveness of the DIFL approach, and thus, in order to achieve peak DIFL model performance, the disease presentation of the datasets should be analyzed whenever possible before deciding on the source and target datasets.

Chapter 6: Conclusion

It is apparent that conventional non-DIFL models are unable to generalize to domains on which they were not trained on, even if the actual classification task is identical. This is largely due to the presence of *domain shift*, which causes a non-DIFL model trained on a source dataset to underperform when tested on other target datasets. Utilizing a DIFL approach mitigates this problem by using unlabeled data from the target dataset to produce more generalized features from the source and target dataset, upon which the classification task is then conducted. As such, the DIFL approach enables us to perform classification tasks on the target dataset, with a much better performance than standard non-DIFL deep learning algorithms, even when the data from the target dataset is unlabeled.

It is observed that the DIFL model performs significantly better than non-DIFL models. Additionally, there are circumstances in which the DIFL model has the potential to perform equivalently as the non-DIFL models which are directly trained on the target dataset. However, achieving such ideal results is largely dependent on the disease presentation among the datasets that are utilized, and choosing the optimal source and target dataset is key in achieving optimal DIFL model performance.

In the context of specific datasets that were used in this research, it is observed that the DIFL model achieves relatively good performance when the US or India dataset is used as the source dataset, and the China dataset is used as the target dataset. When these DIFL models were tested on the target dataset (China), they achieved results that were almost comparable to that of standard non-DIFL algorithms trained directly on the China dataset. However, while these two combinations produced near-ideal results, all DIFL models significantly outperform their non-DIFL counterparts when trained on the source dataset and tested on the target dataset.

Chapter 7: Future Work

The field of DIFL, and domain adaptation in general, has a huge potential for many purposes, and further refinement would enable it to be applied to a large number of other purposes. This section discusses a few of the possible extensions through which one could continue the work that has been done in this paper.

The main facet of this experiment which could be improved to achieve better results would be hyperparameter tuning, particularly that of the classification step learning rate and the domain invariance step learning rate. Deciding the appropriate classification step and domain invariance step learning rates was done by conducting a few trial and error test runs, to decide which set of values would provide the best set of results. However, instead of using trial and error to determine the best set of values for these hyperparameters, one could take a more systematic approach, by using linear algebra to solve for the ideal values of these hyperparameters, by including them along with the mathematical functions of the neural networks. This can be done by accounting for the ideal step sizes within the architecture of the DIFL model, and cleverly manipulating the loss functions that are used. While this might make the DIFL neural network architecture more complex, it would produce

better results as it would ensure that the DIFL model is achieving the right balance of generalization and learning classification features.

Apart from utilizing DIFL algorithms for classification of TB from chest X-rays, one could also attempt to implement similar DIFL algorithms for classification of other diseases, such as lung cancer, pneumonia, and even CoViD-19. Additionally, instead of limiting ourself to chest X-rays, domain adaptation techniques can be implemented on other types of medical imaging modalities, such as mammograms, CT scans and MR images to name a few.

Lastly, another aspect which could be experimented with would be evaluating the effectiveness of the DIFL algorithm when multiple source datasets or multiple target datasets are involved. While it is shown in this paper that the DIFL algorithm, in the context of TB screening from chest X-rays, can generalize and mitigate the domain shift between a single source dataset and a single target dataset, its effectiveness in dealing with multiple source or target datasets is unknown. This would be a relevant extension to the research done in this thesis, as in reality, one may have access to multiple source and target datasets. The DIFL model could be modified slightly to make this change possible, and experiments could be conducted to assess how using multiple source or target datasets affects the end result.

Bibliography

[1] Hal Daume III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of artificial Intelligence research*, 26:101–126, 2006.

[2] Yun Liu, Yu-Huan Wu, Yunfeng Ban, Huifang Wang, and Ming-Ming Cheng. Rethinking computer-aided tuberculosis diagnosis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2646–2655, 2020.

[3] Novi Patricia and Barbara Caputo. Learning to learn, from transfer learning to domain adaptation: A unifying perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1442–1449, 2014.

[4] Mohsen Ghafoorian, Alireza Mehrtash, Tina Kapur, Nico Karssemeijer, Elena Marchiori, Mehran Pesteie, Charles RG Guttmann, Frank-Erik de Leeuw, Clare M Tempany, Bram Van Ginneken, et al. Transfer learning for domain adaptation in mri: Application in brain lesion segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 516–524. Springer, 2017.

[5] Wouter M Kouw and Marco Loog. An introduction to domain adaptation and transfer learning. *arXiv preprint arXiv:1812.11806*, 2018.

[6] Yi Liu, Chao Yang, Kaixin Liu, Bocheng Chen, and Yuan Yao. Domain adaptation transfer learning soft sensor for product quality prediction. *Chemometrics and Intelligent Laboratory Systems*, 192:103813, 2019.

[7] Joey Tianyi Zhou, Ivor W Tsang, Sinno Jialin Pan, and Mingkui Tan. Heterogeneous domain adaptation for multiple classes. In *Artificial intelligence and statistics*, pages 1095–1103. PMLR, 2014.

[8] Jingjing Li, Ke Lu, Zi Huang, Lei Zhu, and Heng Tao Shen. Transfer independently together: A generalized framework for domain adaptation. *IEEE transactions on cybernetics*, 49(6):2144–2155, 2018.

[9] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *arXiv preprint arXiv:1805.12018*, 2018.

[10] Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Holger Roth, Andriy Myronenko, Daguang Xu, and Ziyue Xu. When unseen domain generalization is unnecessary? rethinking data augmentation. *arXiv preprint arXiv:1906.03347*, 2019.

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

[12] Avishek Saha, Piyush Rai, Hal Daumé, Suresh Venkatasubramanian, and Scott L DuVall. Active supervised domain adaptation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 97–112. Springer, 2011.

[13] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5715–5725, 2017.

[14] Hal Daumé III, Abhishek Kumar, and Avishek Saha. Frustratingly easy semi-supervised domain adaptation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 53–59, 2010.

[15] Ting Yao, Yingwei Pan, Chong-Wah Ngo, Houqiang Li, and Tao Mei. Semi-supervised domain adaptation with subspace learning for visual recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2142–2150, 2015.

[16] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8050–8058, 2019.

[17] Mahsa Baktashmotlagh, Mehrtash T Harandi, Brian C Lovell, and Mathieu Salzmann. Unsupervised domain adaptation by domain invariant projection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 769–776, 2013.

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

[19] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. *arXiv preprint arXiv:1602.04433*, 2016.

[20] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017.

[21] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4500–4509, 2018.

[22] Sumit Chopra, Suhrid Balakrishnan, and Raghuraman Gopalan. Dlid: Deep learning for domain adaptation by interpolating between domains. In *ICML workshop on challenges in representation learning*, volume 2. Citeseer, 2013.

[23] Jinkun Cao, Hongyang Tang, Hao-Shu Fang, Xiaoyong Shen, Cewu Lu, and Yu-Wing Tai. Cross-domain adaptation for animal pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9498–9507, 2019.

[24] Jieli Zhou, Baoyu Jing, Zeya Wang, Hongyi Xin, and Hanghang Tong. Soda: Detecting covid-19 in chest x-rays with semi-supervised open set domain adaptation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2021.

[25] Ali Madani, Mehdi Moradi, Alexandros Karargyris, and Tanveer Syeda-Mahmood. Semi-supervised learning with generative adversarial networks for chest x-ray classification with ability of data domain adaptation. In *2018 IEEE 15th International symposium on biomedical imaging (ISBI 2018)*, pages 1038–1042. IEEE, 2018.

[26] Rahul Venkataramani, Hariharan Ravishankar, and Saihareesh Anamandra. Towards continuous domain adaptation for medical imaging. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 443–446. IEEE, 2019.

[27] Cheng Chen, Qi Dou, Hao Chen, and Pheng-Ann Heng. Semantic-aware generative adversarial nets for unsupervised domain adaptation in chest x-ray segmentation. In *International workshop on machine learning in medical imaging*, pages 143–151. Springer, 2018.

[28] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng-Ann Heng. Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 865–872, 2019.

[29] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[30] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475, 2014.

[31] Arun Chauhan, Devesh Chauhan, and Chittaranjan Rout. Role of gist and phog features in computer-aided diagnosis of tuberculosis without segmentation. *PloS one*, 9(11):e112980, 2014.