XDT-CXR: Investigating Cross-Disease Transferability in Zero-Shot Binary Classification of Chest X-Rays

Umaima Rahman UMAIMA.RAHMAN@MBZUAI.AC.AE
Abhishek Basu ABHISHEK.BASU@MBZUAI.AC.AE
Muhammad Uzair Khattak UZAIR.KHATTAK@MBZUAI.AC.AE
Department of Computer Vision
Mohamed Bin Zayed University of Artificial Intelligence
Masdar City, Abu Dhabi, UAE.

Aniq Ur Rahman ANIQ.RAHMAN@ENG.OX.AC.UK
Department of Engineering Science
University of Oxford
Oxford, UK.

Abstract

This study explores the concept of cross-disease transferability (XDT) in medical imaging, focusing on the potential of binary classifiers trained on one disease to perform zero-shot classification on another disease affecting the same organ. Utilizing chest X-rays (CXR) as the primary modality, we investigate whether a model trained on one pulmonary disease can make predictions about another novel pulmonary disease, a scenario with significant implications for medical settings with limited data on emerging diseases. The XDT framework leverages the embedding space of a vision encoder, which, through kernel transformation, aids in distinguishing between diseased and non-diseased classes in the latent space. This capability is especially beneficial in resource-limited environments or in regions with low prevalence of certain diseases, where conventional diagnostic practices may fail. However, the XDT framework is currently limited to binary classification, determining only the presence or absence of a disease rather than differentiating among multiple diseases. This limitation underscores the supplementary role of XDT to traditional diagnostic tests in clinical settings. Furthermore, results show that XDT-CXR as a framework is able to make better predictions compared to other zero-shot learning (ZSL) baselines. Our code and pre-trained models are available at github.com/rumaima/xdt_cxr.

1. Introduction

Motivated by the fact that diseases on the same organ have visual similarities (Özger et al., 2020), we aim to answer the question: *"Can a binary classifier be trained on one disease to perform zero-shot classification on another disease on the same organ using medical images of the same modality?"* We refer to the capability of a model to perform cross-disease zero-shot classification as cross-disease transferability (XDT). In this work, we focus on chest X-rays (CXR) to construct a cross-disease transferable binary classifier. For example, if we have a model that is trained on pneumonia dataset (Kermany et al., 2018) but it is able to accurately classify COVID-19 samples (Rahman et al., 2021) then this will be helpful for

cases where we have limited training data on COVID. Furthermore, in the future if we have a novel disease that affects the same organ then we can make predictions leveraging the cross-disease capabilities of that model. Furthermore, XDT will be particularly beneficial for resource-limited settings in under-developed countries with scant laboratory facilities. Also, in low TB burden countries where TB diagnosis may not be as readily discernible on X-rays and therefore would benefit from the application of cross-disease transferability. Moreover, the XDT framework can also be applied to train on a smaller labeled dataset and evaluate on a larger unlabeled dataset. The XDT framework heavily relies on the success of the embedding space of the vision encoder which is further transformed through a kernel to enable segregation among classes in the latent space.

Our XDT framework has a limitation that it can only perform binary classification i.e., it can say whether a medical image is diagnosed with a disease or not and it cannot be used to classify an image among different diseases. This is supported by the fact that visual examination like X-rays and CT-scans especially for lungs is performed in addition to viral/bacterial tests and only serves as a reinforcement. Furthermore, this study highlights several significant insights into the application of our XDT framework, particularly regarding novel disease detection and resource management. One of the most notable findings is the ability to use models trained on one disease, such as pneumonia, to predict another disease, like COVID-19. This cross-disease transferability is particularly valuable for making quick and basic decisions in triage scenarios. For instance, clinicians can use XDT to categorize two breathless patients into different sickness categories, helping to determine which patient is more critically ill based on X-ray results. The rapid and low-radiation nature of X-rays makes them an efficient tool for initial assessments. This is especially beneficial in rural or resource-limited settings where clinicians must decide which patients need to be sent to better-equipped urban hospitals. By serving as a peripheral screening tool, XDT can assist in managing resources effectively within budget constraints. In terms of pathology testing, machine learning models can aid in answering specific clinical questions, such as whether a patient has COVID-19. While traditional models might identify abnormal respiratory patterns without specifying the exact condition, an effective XDT model can act as a rule-out test. If the model indicates that a patient is healthy, it provides reassurance to discharge them, reducing unnecessary hospital admissions. Additionally, XDT can help identify other pathologies that mimic respiratory diseases. For example, in cases of pulmonary embolism, where a patient feels like they have pneumonia but actually has a blood clot in the lungs, the model can differentiate between the conditions, aiding in accurate diagnosis and appropriate treatment.

Generalizable Insights about Machine Learning in the Context of Healthcare

This study explores the concept of cross-disease transferability in medical imaging, focusing on the potential of binary classifiers trained on one disease to perform zero-shot classification on another disease affecting the same organ. XDT capability is especially beneficial in resource-limited environments or in regions with low prevalence of certain diseases, where conventional diagnostic practices may fail. XDT can serve as a peripheral screening tool, helping clinicians categorize patients, answer specific diagnostic questions, and rule out diseases, thereby optimizing patient care and resource allocation. Furthermore, while tra-

ditional models might identify abnormal respiratory patterns without specifying the exact condition, an effective XDT model can act as a rule-out test. If the model indicates that a patient is healthy, it provides reassurance to discharge them, reducing unnecessary hospital admissions. Although currently limited to binary classification, determining only the presence or absence of a disease, the XDT framework demonstrates superior performance compared to other zero-shot learning (ZSL) baselines. This underscores its supplementary role to traditional diagnostic tests and highlights its potential to enhance diagnostic accuracy, optimize healthcare delivery, and support clinical decision-making in various settings.

2. Related Work

2.1. Zero Shot Learning

Zero-Shot Learning (ZSL) is a machine learning technique that addresses the challenge of classifying previously unseen classes during the training process. The key idea behind ZSL is to leverage the semantic information of the seen classes to generalize and transfer knowledge to the unseen classes. Hayat et al. (2021b) addresses a fundamental limitation of supervised learning models in chest X-ray (CXR) classification - their inability to predict unseen disease classes during inference. They propose a multi-label generalized zero-shot learning (CXR-ML-GZSL) network to overcome this. The innovation of CXR-ML-GZSL is its ability to learn a visual representation guided by the input's corresponding semantics extracted from medical text. This allows the model to map visual and semantic modalities to a shared latent space, ensuring relevant labels are ranked higher. Crucially, the network is trained only on seen classes, with no auxiliary data for unseen classes, demonstrating its generalisation capability. Experiments on the NIH Chest X-ray dataset show that CXR-ML-GZSL outperforms baselines in recall, precision, F1 score, and ROC-AUC. Paul et al. (2021) proposed a novel strategy for generalized zero-shot diagnosis of chest radiographs. They leverage the potential of multi-view semantic embedding, which is a promising yet underexplored approach for zero-shot learning (ZSL). Their design also incorporates a self-training phase to address the issue of noisy labels and improve performance on classes not seen during training. Through rigorous experiments, they demonstrate that their model, trained on a single dataset, can consistently perform well across test datasets from diverse sources, including those with significantly different quality.

2.2. Deep Learning for Chest X-Rays

Convolutional Neural Networks (CNNs) have emerged as a popular deep learning approach in the domain of medical image classification. This can be attributed to several key characteristics of CNNs, including their capacity to effectively learn complex features from data using a relatively smaller number of trainable parameters (Nour et al., 2020). Additionally, the weight-sharing mechanism employed in CNNs allows for efficient utilization of the available parameters, leading to improved performance and generalization capabilities (LeCun et al., 2015), which are crucial in medical applications where labeled data can be scarce.

Al-Waisy et al. (2023) developed the COVID-CheXNet, a hybrid deep learning framework for timely diagnosis of COVID-19 infection using chest X-ray images, in response to the increasing pressure on healthcare systems during the COVID-19 outbreak. The system

employed a multi-step approach, first enhancing the contrast of the X-Ray image through contrast-limited adaptive histogram equalization, the noise level was reduced through Butterworth bandpass filtering, and then fusing the results obtained from two pretrained deep learning models, ResNet34 (He et al., 2016) and High-Resolution Network, using a parallel architecture to provide radiologists with high confidence in discriminating between healthy and COVID-19-infected individuals. Kundu et al. (2021) developed a computer-aided diagnosis system for automatic pneumonia detection using chest X-ray images. The authors employed deep transfer learning to address the limited availability of data and designed an ensemble of three CNN models, including GoogLeNet (Szegedy et al., 2015), ResNet-18, and DenseNet-121 (Huang et al., 2017). Their novel approach involved using a weighted average ensemble technique, where the weights assigned to the base learners were determined by fusing the scores of four standard evaluation metrics: precision, recall, F1-score, and the area under the curve, rather than relying on experimental tuning. However, in some instances the ensemble framework failed to produce correct predictions, and because three CNN models are required to train the proposed ensemble, the computation cost is higher than that of the CNN baselines in the literature.

2.3. Manifold Learning

Manifold learning is a family of unsupervised techniques for extracting low-dimensional representations from high-dimensional data by exploiting the intrinsic geometry of the data. Prominent manifold learning algorithms include Isometric Feature Mapping (Isomap), Local Linear Embedding (LLE) (Roweis and Saul, 2000), and t-Distributed Stochastic Neighbor Embedding (t-SNE)(Van der Maaten and Hinton, 2008). Souvenir et al. (2006) parametrized cardiopulmonary image sets using Isomap and reorder the images based on these learned parameters. This reordering results in minimal motion between neighboring images, compared to the original temporal ordering. This simplifies the point correspondence problem and allows pairwise deformations to be estimated and extended into global deformation models.

Zhang et al. (2005) and Zhang et al. (2006) build upon active contour frameworks for image segmentation, applying them to noisy cardiopulmonary images. They leverage Isomap with domain-specific distance metrics to learn a parametrization capturing the underlying degrees of freedom in the data. The authors then calculate the contours across all images simultaneously, using the learned parameters as additional shape constraints.

2.4. Vision Language Models

The recent advancements in the field of vision-language models (VLMs) have gained significant attention (Radford et al., 2021b), (Yao et al., 2021). VLMs are pretrained on a vast collection of image-text pairs available on the internet. This extensive pretraining enables the VLMs to learn the underlying relationships between visual and textual information. The pretraining process of VLMs is typically guided by specific vision-language objectives ((Radford et al., 2021a), (Yu et al., 2022), (Yao et al., 2021)), which facilitate the learning of meaningful image-text correspondences from the large-scale dataset.

Vision-language models like CLIP(Radford et al., 2021b) employs a contrastive learning approach between images and text. The objective is to pull the paired image and text

representations closer together in the shared embedding space, while pushing apart the representations of unrelated image-text pairs. Through this contrastive training on large-scale image-text data, the pretrained CLIP model is able to capture rich correspondences between visual and linguistic information. As a result, the pretrained CLIP model can directly perform zero-shot predictions on various tasks by matching the embeddings of any given image and text (e.g., class names), without the need for additional fine-tuning. This zero-shot inference performance also depends on the choice of text prompts used to define the class names. The text prompts for classification can be manually designed using prompt engineering (An et al., 2023) or automatically learned using prompt optimization techniques (Zhou et al., 2022b; Khattak et al., 2023a, 2024, 2023b; Shu et al., 2022; Abdul Samadh et al., 2024; Zhou et al., 2022a).

However, when it comes to specialized domains like healthcare, the applicability of general vision-language models faces limitations. Medical image-text datasets are significantly smaller compared to the vast web-scraped image-caption pairs used to train models like CLIP. Additionally, the semantic relationships between medical images and their corresponding reports can be more nuanced, with potential for false negatives cases where images and reports from different patients convey similar clinical information, yet are incorrectly treated as unrelated during training. To address these challenges, recent research has explored domain-specific adaptations of vision-language models for the medical field. One such approach is MedCLIP (Wang et al., 2022) , which builds upon the core contrastive learning framework of CLIP but introduces several key innovations. First, MedCLIP decouples the image and text modalities, allowing it to effectively scale up the usable training data in a combinatorial manner at a low cost. Second, MedCLIP replaces the standard InfoNCE loss with a semantic matching loss based on medical domain knowledge, helping to eliminate the problematic false negatives encountered in previous methods.

3. Methodology

A medical image classifier (MIC) architecture is depicted in Fig. 1 wherein a medical image $\mathbf{x}$ is passed through an encoder f which generates a latent vector $\mathbf{z} \in \mathbb{R}^d$. This latent vector is then passed through a classifier g to get the normalized class prediction vector $\hat{\mathbf{y}}$.

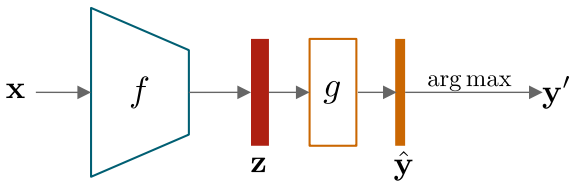

Figure 1: Architecture of the Medical Image Classifier. The encoder f generates a latent vector $\mathbf{z}$, which is passed to the classifier g to produce the normalized class prediction vector $\hat{\mathbf{y}}$. The final output class $\mathbf{y}'$ is obtained via arg max.

Consider a set of pulmonary diseases $\mathfrak{D} \triangleq \{\mathfrak{d}_1, \mathfrak{d}_2, \cdots \mathfrak{d}_n\}$. The binary-class dataset of disease $\mathfrak{d}_i$ is represented by $\mathcal{S}_i \triangleq \mathcal{S}_i^+ \cup \mathcal{S}_i^-$, where $\mathcal{S}_i^+$ denotes the set of $\mathfrak{d}_i$-positive samples labelled $\mathfrak{d}_i$, and $\mathcal{S}_i^-$ is the set of $\mathfrak{d}_i$-negative samples labelled $\mathfrak{d}_i'$. A classifier trained on the

dataset of disease $\mathfrak{d}_i$ is denoted by g_i. Then, the probability that a chest X-ray (CXR) image is classified as $\mathfrak{d}_i$-positive is denoted as $P(\mathfrak{d}_i) \triangleq P(g_i \circ f(\mathbf{x}) = \mathfrak{d}_i : \forall \mathbf{x} \in \mathcal{S}_i)$. Similarly, we also denote the probability of a $\mathfrak{d}_i$-negative samples as $P(\mathfrak{d}_i')$. Consider a classifier g which instead of focusing on a specific disease, classifies whether a given sample from any binary-class CXR dataset is healthy H or unhealthy H' in which a sample is considered healthy if it has none of the diseases in $\mathfrak{D}$.

$$P(\mathsf{H}) = P\left(\bigcap_{i=1}^{n} \mathfrak{d}_i'\right) \implies P(\mathsf{H}) \leq P(\mathfrak{d}_i'), \quad \forall i \in [n]. \tag{1}$$

The binary classifier g_i can be viewed as a hyperplane in $\mathbb{R}^d$ segregating the $\mathfrak{d}_i$ positive and negative samples in $\mathcal{S}_i^+$ and $\mathcal{S}_i^-$ (see Fig. 2). Some negative samples may be positive for some other disease $\mathfrak{d}_j \in \mathfrak{D}, j \neq i$ which can be reflected in the generalised healthy vs. unhealthy classifier g as the relation $P(\mathsf{H}) = P(\mathfrak{d}_i') - \epsilon_i$ which forms the basis of Proposition 1.

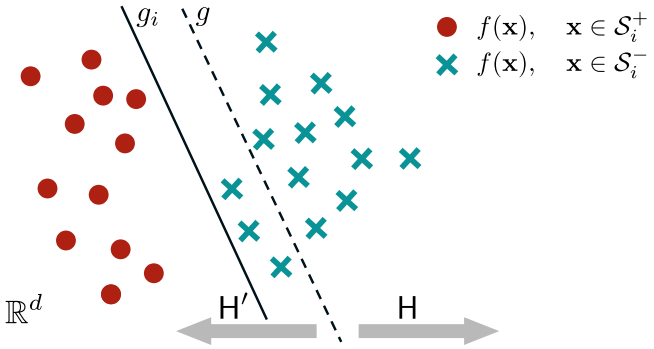

Figure 2: Decision boundaries of disease-specific and healthy/unhealthy binary classifiers. The figure shows two decision boundaries: g_i for a disease-specific classifier and g for a general classifier, separating the feature space $\mathbb{R}^d$ into regions H' and H. Samples $f(\mathbf{x}) \in \mathcal{S}_i^+$ (disease-specific) and $f(\mathbf{x}) \in \mathcal{S}_i^-$ (healthy) are represented by different markers.

Proposition 1 $P(\mathsf{H}) = P(\mathfrak{d}_i') - \epsilon_i$, $\exists \epsilon_i \in [0, P(\mathfrak{d}_i')]$ $\forall i \in [n]$.

It is possible to find a *decision boundary* outlined by the classifier $g^\star$, and an encoding of the images $f^\star(\mathbf{x}) \in \mathbb{R}^d$ such that the *latent representation* of the healthy samples $\mathfrak{d}_i'$ are on one side of the decision boundary, and that of the unhealthy samples $\mathfrak{d}_i$ are on the other. The inter-class segregation as outlined in Fig. 3 will be more feasible with increasing $d \in \mathbb{N}$ as the degrees of freedom for the encoder $f^\star$ and the classifier $g^\star$ increase. We formalise the above idea in Theorem 2.

Theorem 2 *For the diseases* $\mathfrak{D}$*, and datasets* $\mathcal{S}_i$ $\forall i \in [n]$*,* $\exists d \in \mathbb{N}$ *for which* $\exists f^\star, g^\star : \forall i \in [n]$

$$P(\arg\max g^\star \circ f^\star(\mathbf{x}) = \mathfrak{d}_i \mid \mathbf{x} \in \mathcal{S}_i^+) = 1,$$
$$P(\arg\max g^\star \circ f^\star(\mathbf{x}) = \mathfrak{d}_i' \mid \mathbf{x} \in \mathcal{S}_i^-) = 1.$$

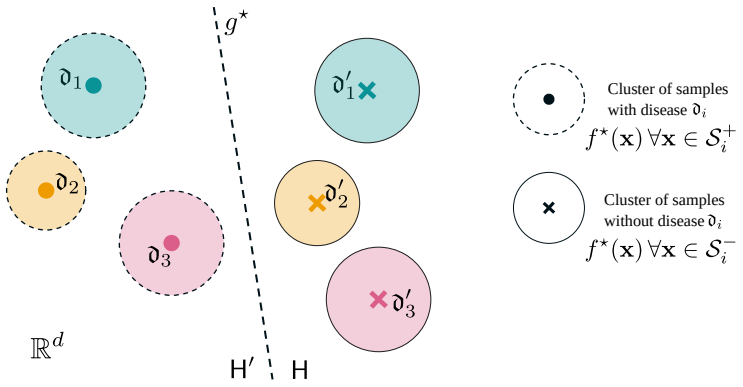

Figure 3: Illustration of a general classifier depicting clusters of samples with and without diseases $\mathfrak{d}_i$ and $\mathfrak{d}'_i$. The clusters of samples with disease $\mathfrak{d}_i$ are enclosed in dashed circles, while the clusters of samples without disease $\mathfrak{d}_i$ are in solid circles. The decision boundary g^* differentiates between samples with and without diseases. The regions H and H' represent the spaces of the samples.

Proof From the consequence of Cover's theorem (Cover, 1965), it is established that a set of datapoints which is not linearly separable in a given dimension, can be made linearly separable with high probability through a non-linear transformation of the datapoints into a higher dimension. The function $f^\star$ which is a neural network acts as the non-linear transformation kernel, which can be affirmed by the universal approximation theorem (Guliyev and Ismailov, 2018) which implies that any function can be approximated by a neural network by having the appropriate weights. ■

Contrary to the setting in Theorem 2, we will be learning an encoder f, and classifier g through supervised training for a disease $\mathfrak{d}_i \in \mathfrak{D}$, and performing zero-shot evaluation using the resultant model on the diseases $\mathfrak{d}_j \in \mathfrak{D} \setminus \{\mathfrak{d}_i\}$.

Corollary 3 *Given disease $\mathfrak{d}_i \in \mathfrak{D}$, and its corresponding dataset $\mathcal{S}_i$, $\exists f_i^\star, g_i^\star$:*

$$
\begin{aligned}
&P(\arg\max g_i^\star \circ f_i^\star(\mathbf{x}) = \mathfrak{d}_i \mid \mathbf{x} \in \mathcal{S}_i^+) = 1, && && \textit{(Supervised)}\\
&P(\arg\max g_i^\star \circ f_i^\star(\mathbf{x}) = \mathfrak{d}'_i \mid \mathbf{x} \in \mathcal{S}_i^-) = 1, && && \textit{(Supervised)}\\
&P(\arg\max g_i^\star \circ f_i^\star(\mathbf{x}) = \mathfrak{d}_j \mid \mathbf{x} \in \mathcal{S}_j^+) \leq 1, && \forall j \in [n] \setminus \{i\}, && \textit{(Zero-Shot)}\\
&P(\arg\max g_i^\star \circ f_i^\star(\mathbf{x}) = \mathfrak{d}'_j \mid \mathbf{x} \in \mathcal{S}_j^-) \leq 1, && \forall j \in [n] \setminus \{i\}. && \textit{(Zero-Shot)}
\end{aligned}
$$

Remark 4 *Through Corollary 3, we show that given the information of only disease $\mathfrak{d}_i$ dataset, it is still theoretically feasible to construct an encoder-classifier pair (f_i, g_i) which matches the performance of a general encoder-classifier pair $(f^\star, g^\star)$. However, even for $(f^\star, g^\star)$ the performance guarantees do not extend beyond the training data. Moreover, finding the encoder-classifier using one disease dataset and testing it on another disease related to the same organ and modality is an interesting research direction to verify if the model can exhibit cross-disease transferability.*

3.1. Problem Formulation

We aim to design a medical image classifier (like Fig. 4) consisting of an encoder-classifier pair (f, g) trained on a labelled dataset $\mathcal{S}_i = \mathcal{S}_i^+ \cup \mathcal{S}_i^-$ of a disease $\mathfrak{d}_i \in \mathfrak{D}$. The learnt model is then used for zero-shot evaluation of the datasets $\mathcal{S}_j, \forall j \in [n] \setminus \{i\}$. Metrics like classification accuracy, and F1 score can be used to judge the performance of the model.

3.2. Network Architecture

We split the encoder f into two parts through $h \circ f_V$, where f_V is the vision encoder of MedCLIP which remains frozen throughout, and h is a trainable transformer resulting in the latent representation $\mathbf{z} \in \mathbb{R}^d$. The latent representation is then passed to a classifier g which returns a normalized class prediction vector $\hat{\mathbf{y}}$. The transformer h, and the classifier g are trained through a supervised objective, wherein the loss function $\mathcal{L}$ is passed $\hat{\mathbf{y}}$, along with the true label $\mathbf{y}$ associated with the sample $\mathbf{x} \in \mathcal{S}$, where $\mathcal{S}$ is the training data.

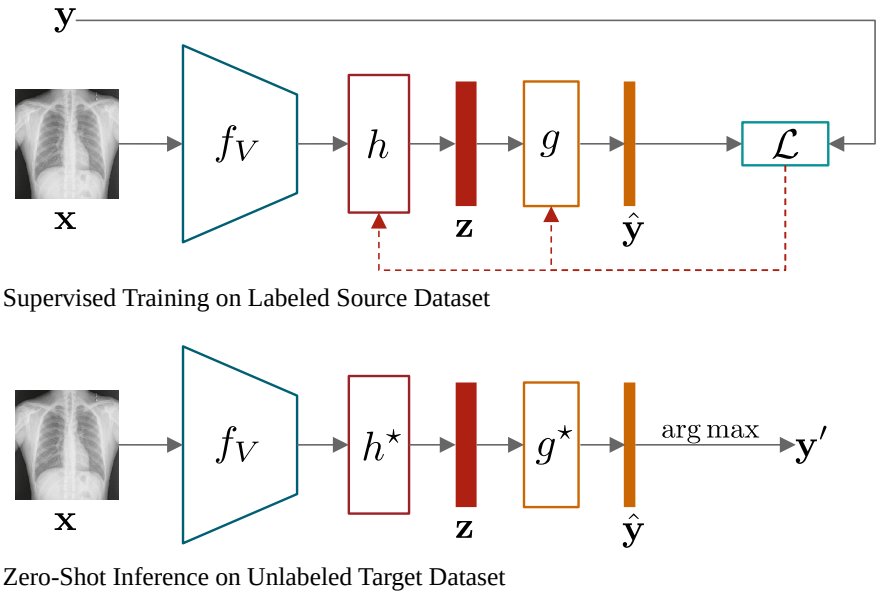

Figure 4: Illustration of the Training and inference networks. The encoder is split into two parts: $h \circ f_V$, where f_V is the frozen vision encoder of MedCLIP, and h is a trainable transformer resulting in the latent representation $\mathbf{z}$. The latent representation $\mathbf{z}$ is passed to a classifier g, which returns a normalized class prediction vector $\hat{\mathbf{y}}$. During supervised training, the transformer h and classifier g are trained using a loss function $\mathcal{L}$, which takes $\hat{\mathbf{y}}$ and the true label $\mathbf{y}$. In zero-shot inference, the trained transformer $h^\star$ and classifier $g^\star$ are used to output the class $\arg\max g^\star \circ h^\star \circ f_V(\mathbf{x})$.

The transformer and classifier resulting from the supervised training is dentoed as $h^\star$, and $g^\star$, respectively. During zero-shot inference, for an input $\mathbf{x}$, we get the output class $\arg\max g^\star \circ h^\star \circ f_V(\mathbf{x})$.

3.3. Loss Function

We need a loss function that brings the latent representations of samples from the same class closer, and pushes apart samples from different classes which should result in clusters that are far apart. As the gap between the clusters of the two classes increases, it gets easier to find a hyperplane to separate them, which can act as a classifier. Therefore, we make use of a logarithmic contrastive loss defined as follows:

Definition 5 *The Logarithmic Contrastive Loss (LC) for dataset $\mathcal{S}$ is defined as:*

$$\mathcal{L}_{\mathrm{LC}}(\mathcal{S}) \triangleq \frac{1}{\binom{|\mathcal{S}|}{2}} \sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{S}} \underbrace{(2\mathbf{y}_i^\top \mathbf{y}_j - 1) \cdot \log\left(\|\mathbf{z}_i - \mathbf{z}_j\|\right)}_{\text{Clustering}} + \lambda \cdot \underbrace{\Big(\mathcal{H}(\hat{\mathbf{y}}_i, \mathbf{y}_i) + \mathcal{H}(\hat{\mathbf{y}}_j, \mathbf{y}_j)\Big)}_{\text{Cross-Entropy}}; \ \lambda \in \mathbb{R}^+,$$

where $\mathbf{y}_i = l(\mathbf{x}_i)$ is the one-hot encoded true label, $\mathbf{z}_i = h \circ f_V(\mathbf{x}_i)$ is the latent representation. The predicted class label is one-hot encoded as $\hat{\mathbf{y}}_i = \mathbf{e}_{\arg\max g \circ h \circ f_V(\mathbf{x}_i)}$ where $\mathbf{e}_k$ is the k^{th} standard basis, and $\mathcal{H}(\cdot, \cdot)$ is the cross-entropy function.

The cross-entropy term ensures that the encoder and classifier learnt during training do not just focus on the clustered appearance in the latent space, but also on the accuracy on the training set. While the Cross-Entropy term is prone to overfitting, the Clustering term acts as a regulariser and generalises the model for unseen diseases, supported by Remark 4.

In the clustering term, we write $\log(\|\mathbf{z}_i - \mathbf{z}_j\|)$ instead of simply $\|\mathbf{z}_i - \mathbf{z}_j\|$ to ensure that the sum over all the pair of datapoints $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{S}$ does not become too high (`Inf`) during computation. The hyperparamter $\lambda \in \mathbb{R}^+$ controls the level of influence the cross-entropy term has on the training objective. We can also replace $\|\mathbf{x}_i - \mathbf{z}_j\|$ with $\|\mathbf{x}_i - \mathbf{z}_j\| + \epsilon$, $\epsilon \approx 0^+$ to prevent the argument of $\log(\cdot)$ from becoming zero.

In this work, we will also be using cross-entropy loss, and euclidean contrastive loss in out framework to compare against the logarithmic contrastive loss defined above.

Definition 6 *The Cross-Entropy Loss (CE) for dataset $\mathcal{S}$ is defined as:*

$$\mathcal{L}_{\mathrm{CE}}(\mathcal{S}) \triangleq \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x}_i \in \mathcal{S}} \mathcal{H}(\hat{\mathbf{y}}_i, \mathbf{y}_i). \tag{2}$$

Definition 7 *The Euclidean Contrastive Loss (EC) for dataset $\mathcal{S}$ is defined as:*

$$\mathcal{L}_{\mathrm{EC}}(\mathcal{S}) \triangleq \frac{1}{\binom{|\mathcal{S}|}{2}} \sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{S}} (2\mathbf{y}_i^\top \mathbf{y}_j - 1) \cdot \|\mathbf{z}_i - \mathbf{z}_j\|. \tag{3}$$

4. Experimental Settings

In this section, we present the design of the experiments conducted to evaluate the performance of our proposed method. The experiments were carefully designed to test the efficacy of our approach across various scenarios and datasets. We begin by detailing the experimental setup, including the datasets used, implementation details and baseline methods for comparison.

4.1. Datasets

We have used four publicly available Chest X-Ray (CXR) datasets in our experiments:

Guangzhou-PN (G) Pneumonia dataset published in Kermany et al. (2018), consists of 5,232 pediatric CXR images, which were collected from patients aged one to five years old at Guangzhou Women and Children's Medical Center in Guangzhou, Guangdong Province, China. There are 1,349 normal and 3,883 Pneumonia x-rays (2,538 x-rays of Bacterial pneumonia and 1,345 Viral pneumonia).

Montgomery-TB (M) Montgomery dataset mentioned in Jaeger et al. (2014), is a collection of 138 posterior-anterior CXR. Which was collected as part of the tuberculosis control program by the Department of Health and Human Services of Montgomery County, MD, USA. The collection includes 80 normal and 58 abnormal x-rays indicating manifestations of tuberculosis. The normal x-rays do not exhibit any signs of tuberculosis or other abnormalities.

Shenzhen-TB (S) Shenzhen Hospital dataset also in Jaeger et al. (2014), we will refer to as Shenzhen-TB consists of 662 CXR images as part of the routine care at the No.3 Hospital in Shenzhen, Guangdong Province, China. Of these images 326 normal and 336 abnormal x-rays.

Covid (C) Covid dataset from Rahman et al. (2021) use 1383 test CXR images, this is the largest public COVID positive database. We divided this set into equal number of images in each class, positive (COVID) and negative (non-COVID).

4.2. Baselines

Our work Cross-Disease Transferability for Chest X-Rays (XDT-CXR) suggests that maximising the cluster gap among the diseases can help in making the model (trained on one dataset) transferable to other datasets. XDT-CXR is based on the vision encoder of MedCLIP (Wang et al., 2022) and a vision transformer that helps to learn the clusters in the latent space.

Statistical Best We consider a theoretical model which consistently predicts the same class for all samples. If the number of samples of one class is more than the other in the dataset in question, the statistical best will always predict the majority label. The Statistical Best model serves an an absolute benchmark as any model which is subpar to it has not learnt anything meaningful from the data.

MedCLIP MedCLIP (Wang et al., 2022) is a Vision Language model that trains CLIP Radford et al. (2021b) on medical data. We perform a zero-shot evaluation on our dataset.

LaFter This is an approach with an unsupervised tuning of a zero-shot classifier leveraging the cross-modal transferabilities of VLM Mirza et al. (2024)

TPT Test-Time Prompt Tuning (TPT) Shu et al. (2022) utilises the embedding space created by a VLM and adapts a prompt at test-time relying on the most confident predictions of multiple augmentations of the test image.

CXR-ML-GZSL This is a Multi-label Generalised ZSL on the NIH Chest X-Ray dataset which consists of multiple diseases. In this work the model learns to map visual and semantic modalities to a shared latent space, ensuring relevant labels are ranked higher. It is named ML-GZSL due to the reason of it being trained only on seen classes without any auxiliary data for unseen classes. CXR-ML-GZSL Hayat et al. (2021a) uses `DenseNet` as the vision encoder. For all the other baselines we have used MedCLIP's vision encoder as the backbone.

4.3. Metrics

For our Cross-dataset (XD) zero-shot (ZS) evaluation for binary classification problem, we use accuracy and F1-score to predict the model's capability.

Accuracy Accuracy is a metric commonly used in classification tasks to quantify the proportion of correctly classified instances among the total number of instances evaluated.

$$\texttt{Acc} \triangleq \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{4}$$

where TP denotes true positives, TN denotes true negatives, FP denotes false positives, and FN denotes false negatives.

F1 score The F1-score is a harmonic mean of precision and recall, providing a balance between these two metrics. It is particularly useful when dealing with imbalanced datasets.

$$\texttt{F1} \triangleq \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}, \tag{5}$$

where precision is the ratio of true positive predictions to the total number of positive predictions, and recall is the ratio of true positive predictions to the total number of actual positive instances.

Relative Accuracy The relative change of accuracy with respect to the statistical best accuracy is defined as the relative accuracy

$$\texttt{Acc'} \triangleq \frac{\texttt{Acc} - \texttt{Acc}_{\text{SB}}}{\texttt{Acc}_{\text{SB}}}, \tag{6}$$

where $\texttt{Acc}_{SB}$ is the statistical best accuracy and `Acc` is the cross-disease zero-shot accuracy of our model.

4.4. Implementation details

This section provides an in-depth overview of the implementation details of our proposed method to ensure reproducibility and clarity of our approach.

4.4.1. Supervised Training on Labeled Source Dataset

In this stage we input an image which passes through a a frozen vision encoder from MedCLIP and the visual embeddings thus obtained are then fed into a trainable transformer. We used this pipeline to learn class-wise separable features, as theoretically motivated in Section 3. Our model processes visual data using a multi-layer transformer encoder followed by a

linear classification layer. Specifically, it consists of a `nn.TransformerEncoder` with four `nn.TransformerEncoderLayer` layers, each with an input and output dimension of 512, four attention heads, and a feedforward network dimension of 256. The transformed input features are then passed through a `nn.Linear` layer, reducing the dimensionality from 512 to 16 for downstream task of classification. The choice of using a transformer compared to other architectures was motivated by the improved accuracy by using transformers as discussed in Appendix A. During supervised training, both the transformer and the classifier are updated using a loss function that compares the predicted class vector with the true label.

4.4.2. Zero-Shot Inference on Unlabeled Target Dataset

For zero-shot inference, the trained transformer and classifier are employed alongside the frozen vision encoder to make predictions on new images. In this phase, the input image is first processed by the frozen vision encoder to extract visual features. These features are then passed through the pre-trained transformer, which transforms them into a latent representation. The classifier takes this latent representation and outputs a prediction vector. The class with the highest score in this prediction vector is selected as the final output. This allows the system to accurately classify images it has not seen during training, leveraging the robust feature extraction of the frozen vision encoder combined with the learned representations from the transformer and classifier.

4.4.3. Generating Baseline Results

Baselines using the LaFter and TPT pipelines were established with their official codebases. For LaFter, GPT-3.5 generated textual descriptions for medical dataset classes with specific chest X-ray prompts, which were encoded into embeddings to train a text classifier using cross-entropy loss. The trained classifier generated pseudo labels for self-supervised learning, and MedCLIP's image encoder replaced CLIP's, using label smoothing cross-entropy loss. Performance was evaluated on the test set for each dataset in Section 4.1, with Table 2 reporting LaFTer's diagonal values as presented in Table 8 of Appendix C. We used the original TPT baseline settings, encoding medical text prompts and class labels into textual embeddings, and augmenting each test image into multiple views encoded into visual embeddings with a MedCLIP backbone. Confidence scores were calculated, unreliable ones filtered out, and remaining scores averaged for final classification. The model minimized entropy in the predicted distribution through backpropagation to fine-tune input prompts, accepting or rejecting them based on their effectiveness in reducing classification uncertainty at test-time. These steps were repeated for each dataset in Section 4.1.

4.5. Model Training

As discussed in section 3.2, our framework first performs supervised training on labeled source dataset and then conducts a zero-shot inference on the unlabeled target dataset. During training, the medical images are passed through the MedCLIP vision encoder and the vision embeddings obtained are send as an input to the transformer. The loss function is used to maximise the cluster distance between the latent representations learned using the vision transformer. Once the model is trained on one dataset, it is then evaluated on a

different dataset. The vision embeddings obtained using a VLM holds valuable information as a result we experimented with CLIP vision encoder with ViT-B/32 backbone as well as MedCLIP vision encoder with Swin Transformer backbone. During exepriments we observed that the results on MedCLIP backbone outperformed the CLIP backbone as discussed in Appendix B. As a result we used MedCLIP backbone for all our experiments. Each dataset was split into a 60-20-20 regime where 60% of the data was used for training, 20% for validation and 20% for testing. The choice of optimizer was critical, and when experiments were conducted using Adam and SGD, we observed that SGD gave better results and hence we used SGD optimiser with a learning rate of 10^{-2}. The experiments were conducted on a 2nd Gen AMD Epyc processors and on a single Nvidia A100 Tensor Core GPU.

5. Results

In this section, we present the outcomes of the experiments conducted to evaluate the performance of our proposed method XDT-CXR.

5.1. Supervised Learning and Zero-Shot Image Classification

In our supervised setting, as discussed in section 3.2 that our model was trained in a supervised manner on a dataset D_i and then evaluated on the test set of the same dataset D_i. The performance evaluation, detailed in Table 1, covers four datasets: TB cases from Shenzhen and Montgomery, COVID-19 cases, and pneumonia cases from Guangzhou, with metrics including accuracy (Acc) and F1-score (F1). For the Shenzhen-TB dataset, the model achieved an accuracy of 73.68% and an F1-score of 0.67, indicating moderate precision and recall. The model performed best on the COVID-19 dataset, with an accuracy of 88.94% and an F1-score of 0.74, highlighting its effectiveness in identifying COVID-19 cases. Conversely, the performance on the Guangzhou-PN dataset was lower, with an accuracy of 58.50%, though the F1-score of 0.69 suggests a reasonable balance of precision and recall. The Montgomery-TB dataset results were strong, with an accuracy of 82.76% and an F1-score of 0.78, reflecting robust performance in TB classification. Furthermore to understand

Table 1: Performance evaluation on the test set of the same dataset. Each dataset represents a chest disease, including TB cases from Shenzhen and Montgomery, COVID-19 cases, and pneumonia cases from Guangzhou. The metrics include accuracy (`Acc`) and F1-score (`F1`).

Shenzhen-TB		Covid		Guangzhou-PN		Montgomery-TB	
`Acc`	`F1`	`Acc`	`F1`	`Acc`	`F1`	`Acc`	`F1`
73.68	0.67	88.94	0.74	58.50	0.69	82.76	0.78

the zero-shot capability, the model trained on dataset D_i is then evaluated on the remaining D_{n-i} datasets in a zero-shot manner. Table 2 gives a comparative analysis of zero-shot evaluation of XDT-CXR on the datasets mentioned in section 4.1 as well as the baselines described in section 4.2. The table highlights the exceptional zero-shot performance of XDT-CXR across various medical image classification tasks. XDT-CXR achieves impressive results even when trained on a different dataset than the one being tested. For instance,

when trained on the Shenzhen-TB dataset, it shows an accuracy of 80.12% and an F1-score of 0.65 on the same dataset, while performing notably well on the COVID-19 dataset with an accuracy of 65.41% and an F1-score of 0.43. Its performance on the Guangzhou-PN dataset, where it achieves an accuracy of 79.31% and an F1-score of 0.77, and on the Montgomery-TB dataset, where it achieves an accuracy of 71.43 and an F1-score of 0.56, further underscores its robust generalization capabilities. The varying performance across different datasets also underscores the importance of the vision encoder's embedding space and the kernel transformation process. Effective segregation in the latent space is crucial for the model to distinguish between normal and pathological features across different diseases. Furthermore, the comparison with other methods, such as MedCLIP and TPT, reveals the competitive nature of the XDT-CXR in the zero-shot scenario, often outperforming these established benchmarks. This demonstrates the potential of the XDT framework to be a valuable tool in settings where rapid adaptation to new diseases is necessary, such as during outbreaks of novel pathogens.

Table 2: Zero-shot evaluation on the test set of a different dataset.

	Shenzhen-TB		Covid		Guangzhou-PN		Montgomery-TB	
Method	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Statistical Best	57.14	0.60	73.75	0.42	62.50	0.77	51.72	0.65
MedCLIP	42.11	0.57	80.84	0.63	80.93	0.86	55.17	0.61
LaFter	57.14	0.00	26.25	0.42	62.5	0.77	51.72	0.00
TPT	50.76	**0.65**	26.19	0.00	72.97	0.58	42.03	1.00
CXR-ML-GZSL	57.14	0.6	73.75	0.42	62.66	0.77	58.62	0.68
XDT-CXR (Shenzhen-TB)	x	x	80.12	0.65	76.28	0.83	75.86	0.72
XDT-CXR (Covid)	65.41	0.43	x	x	**88.14**	0.90	55.17	0.13
XDT-CXR (Guangzhou-PN)	56.39	0.60	52.50	0.44	x	x	**79.31**	0.77
XDT-CXR (Montgomery-TB)	**71.43**	0.56	**81.56**	0.61	79.01	0.82	x	x

Fig. 5 visualizes the performance of various models across the four datasets: Shenzhen-TB, Covid, Guangzhou-PN, and Montgomery-TB. The models compared are Statistical Best, Zero-shot MedCLIP, LaFter, TPT, CXR-ML-GZSL, and XDT-CXR. Each axis represents one of the datasets, with values indicating the accuracy percentages achieved by each model. The observations include XDT-CXR's superior performance and highest accuracy across all datasets. It outperforms all other models in the Covid dataset and maintains strong performance in the Guangzhou-PN and Montgomery-TB datasets. Statistical Best and CXR-ML-GZSL also perform well, whereas LaFter and TPT show lower accuracy. The plot underscores XDT-CXR's robustness and generalization capabilities, highlighting its superior zero-shot performance in diverse medical imaging tasks.

5.2. Impact of Different Loss Functions

We study the significance of our loss function contributing to the zero-shot performance of our framework XDT-CXR. Table 3 and Table 4 highlights the results of the different loss functions compared to the one we used in our training pipeline. The $\mathcal{L}_{\text{LC}}$ loss function

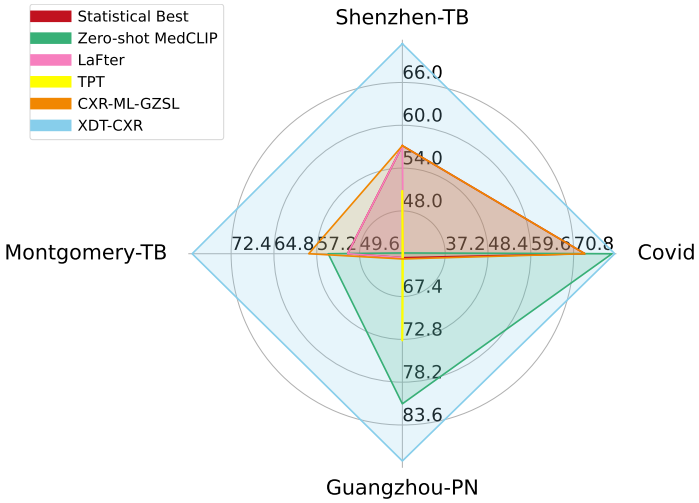

Figure 5: Radar plot showing the accuracy of different baselines across four datasets: Montgomery-TB, Shenzhen-TB, Guangzhou-PN, and Covid. The baselines compared include Statistical Best, Zero-shot MedCLIP, LaFter, TPT, CXR-ML-GZSL, and **XDT-CXR**.

seems to facilitate better generalization of learned features across different diseases, which is crucial for zero-shot learning scenarios. This is likely due to its ability to balance between multiple objectives, enhancing the model's capability to identify universal patterns indicative of various chest diseases. The robustness provided by $\mathcal{L}_{\text{LC}}$ indicates its potential as a reliable choice for training diagnostic models that need to operate across varied and unforeseen medical conditions, reflecting real-world clinical needs where rapid adaptation to new diseases is often necessary. Table 3 presents a comparative analysis of various loss functions used in the training pipeline XDT-CXR across the four datasets. The results show that $\mathcal{L}_{\text{EC}}$ generally performs well, with the highest accuracy on Shenzhen-TB (77.44%) and Guangzhou-PN (72.76%), though its F1 scores are not the highest across all datasets. $\mathcal{L}_{\text{CE}}$ shows variability in performance, with its best F1 score (0.77) achieved on Guangzhou-PN. On the other hand, $\mathcal{L}_{\text{LC}}$, especially with $\lambda = 0.001$, yields consistently high F1 scores and competitive accuracies, notably achieving the highest F1 scores on Covid (0.74) and Montgomery-TB (0.78). Overall, $\mathcal{L}_{\text{LC}}$ with $\lambda = 0.001$ appears to offer balanced performance across different datasets, excelling in F1 scores. Further analysis on the choice of a suitable loss function is available in Appendix D.

Table 3: Comparative analysis of different loss functions for the training pipeline XDT-CXR on the test set of the same dataset

XDT-CXR	Shenzhen-TB		Covid		Guangzhou-PN		Montgomery-TB	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
$\mathcal{L}_{\text{EC}}$	**77.44**	0.66	77.44	0.65	**72.76**	0.82	75.86	0.72
$\mathcal{L}_{\text{CE}}$	42.86	0.60	73.75	0.0	62.50	0.77	48.28	0.65
$\mathcal{L}_{\text{LC}}, \lambda = 0$	75.94	0.69	86.12	0.67	60.26	0.71	**82.76**	0.78
$\mathcal{L}_{\text{LC}}, \lambda = 0.001$	73.68	0.67	**88.94**	0.74	58.50	0.69	**82.76**	0.78

Table 4: Comparative analysis of different loss functions for the training pipeline XDT-CXR on the zero-shot evaluation on the test set of different datasets

XDT-CXR	Loss	Shenzhen-TB		Covid		Guangzhou-PN		Montgomery-TB	
Trained on		Acc	F1	Acc	F1	Acc	F1	Acc	F1
Shenzhen-TB	$\mathcal{L}_{\text{EC}}$	x	x	**86.62**	0.66	37.50	0.00	**79.31**	0.77
Covid		69.93	0.67	x	x	73.88	0.83	**79.31**	0.77
Guangzhou-PN		45.11	0.53	43.82	0.41	x	x	72.41	0.6
Montgomery-TB		**72.93**	0.67	80.69	0.66	**79.17**	0.86	x	x
Shenzhen-TB	$\mathcal{L}_{\text{CE}}$	x	x	**26.25**	0.42	**62.5**	0.77	**48.28**	0.65
Covid		**57.14**	0.66	x	x	37.5	0.00	51.72	0.00
Guangzhou-PN		42.86	0.60	**26.25**	0.42	x	x	48.28	0.65
Montgomery-TB		42.86	0.60	**26.25**	0.42	62.5	0.77	x	x
Shenzhen-TB	$\mathcal{L}_{\text{LC}}$	x	x	80.41	0.64	75.32	0.82	75.86	0.72
Covid	$\lambda = 0$	**73.68**	0.58	x	x	**80.93**	0.83	**82.76**	0.78
Guangzhou-PN		42.86	0.53	35.00	0.37	x	x	72.41	0.73
Montgomery-TB		71.43	0.56	**81.13**	0.60	79.01	0.83	x	x
Shenzhen-TB	$\mathcal{L}_{\text{LC}}$	x	x	80.12	0.65	76.28	0.83	75.86	0.72
Covid	$\lambda = 0.001$	65.41	0.43	x	x	**88.1**4	0.90	55.17	0.13
Guangzhou-PN		56.39	0.60	52.50	0.44	x	x	**79.31**	0.77
Montgomery-TB		**71.43**	0.56	**81.56**	0.61	79.01	0.82	x	x

5.3. Impact of Training Size on Zero-shot Capability

In Fig. 6, we present the relative accuracy vs. size ratio plot. The relative accuracy is defined as the accuracy of the model with respect to the statistical best (majority label memorisation), and the size ratio is the ratio of the test set size to the train set size. Therefore, a size ratio < 1 means that the model was trained on more data than it was tested on, and a size ratio > 1 means that the model was trained on a small dataset and tested on a larger one. We have depicted size ratio < 1 with red circles, and size ratio > 1 with green squares in the scatter plot. The abbreviations S, C, G, and M denote Shenzhen-TB, Covid, Guangzhou-PN, and Montgomery-TB datasets, respectively. The plot compares the performance of models when trained on one dataset and tested on another, highlighting how changes in train/test set sizes impact model accuracy. It shows that models generally achieve higher accuracy when the train set is larger than the test set. This suggests that the size and diversity of the training data are crucial for improving model performance in cross-dataset evaluations.

6. Discussion

The results presented in Tables 2 and 4 provide important insights into the effectiveness of the Cross-Disease Transferability (XDT) framework when applied to chest X-ray (CXR) analysis. In the **supervised learning** scenario, the models demonstrate varied performance across different diseases and datasets. For instance, the model achieved an accuracy of 88.94% and an F1-score of 0.74 for COVID-19, indicating a strong capability to identify

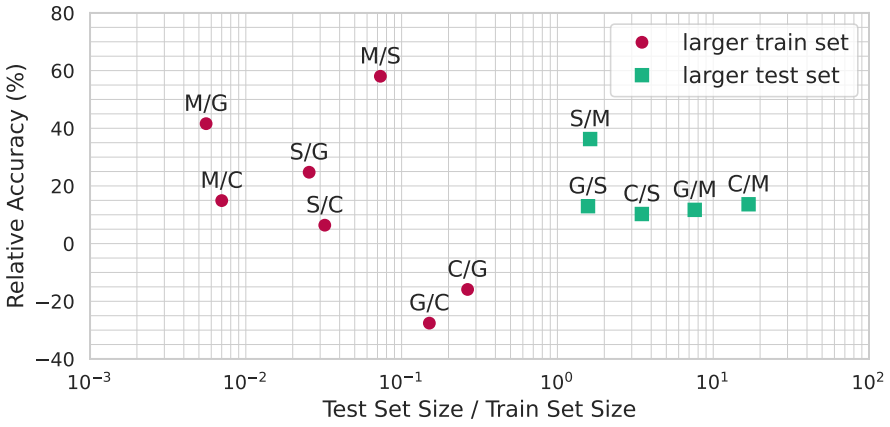

Figure 6: Scatter plot showing the relative accuracy vs. size ratio. A size ratio < 1 means that the model was trained on more data than it was tested on, and a size ratio > 1 means that the model was trained on a small dataset and tested on a larger one.

this disease under controlled, disease-specific training conditions. In contrast, the model's performance on pneumonia from Guangzhou shows a notable drop in accuracy to 58.50%, suggesting that certain visual disease markers may not be as distinctly captured or may overlap with other conditions, complicating the model's learning process. In the **zero-shot classification** scenario, where the models trained on one dataset are tested on others without additional training, the results highlight the potential and challenges of applying the XDT framework to real-world diagnostic tasks. Notably, the XDT-CXR model trained on COVID-19 data performed exceptionally well when applied to the Guangzhou-PN dataset, achieving the highest accuracy of 88.14% and an F1-score of 0.90. This suggests a significant overlap in visual features relevant to the model between these two diseases, or a robustness of the learned features that generalize well across different conditions within the same organ. However, other cross-dataset evaluations, such as the XDT-CXR model trained on the Shenzhen-TB dataset evaluated on COVID-19, showed lower performance (accuracy of 80.12% and F1-score of 0.65) compared to the supervised setting. This indicates the inherent challenge in zero-shot learning where the model must generalize to entirely unseen conditions without losing specificity or sensitivity.

Another noteworthy factor in our experimental setting was that the Guangzhou-PN dataset consists of paediatric data, whereas the other datasets include data from adult patients. In our experiments we investigated two-way cross-disease transferability, i.e., our results show how a model trained on the Guangzhou-PN (G) dataset performs on other datasets in a ZSL setting as well as how other datasets such as Covid (C), Shenzhen-TB (S) or Montgomery-TB (M) which consists of adult CXRs are transferable in giving predictions on Guangzhou-PN datasets as presented in Table 2.

Overall, the findings suggest that while the XDT framework holds significant promise for enhancing diagnostic efficiency using limited training data, extensive testing and refinement are required to maximize its applicability and reliability in diverse clinical settings.

Limitations. The XDT framework has a limitation that it can only perform binary classification i.e., it can say whether a medical image is diagnosed with a disease or not. However, it cannot be used to classify an image among different diseases. This is supported by the fact that visual examination like X-rays and CT-scans especially for lungs is performed in addition to viral/bacterial tests and only serves as a reinforcement. Moreover, despite the promising results as discussed above, the limitations in binary classification capability, where the model can only determine the presence or absence of disease rather than identifying specific conditions, remain a critical challenge. Furtehrmore, we understand the limitation that models trained on lung disease might not very well be transferable to a renal disease or a brain disease, given that it affects different organs. However, cross-modal transferability, i.e., checking if (disease A, modality A) is transferable to (disease B, modality B), is an exciting line of enquiry for the future. Few works in the literature perform cross-modal transferability, such as Gu et al. (2023), which can motivate the work. This limitation points to the necessity for further research into multi-class models within the XDT framework that could handle more complex diagnostic tasks.

7. Conclusion

In conclusion, this study investigates cross-disease transferability (XDT) in medical imaging, showing that binary classifiers trained on one disease can perform zero-shot classification on another affecting the same organ. Using chest X-rays (CXR), we demonstrate that models trained on one pulmonary disease can predict others, which is useful when data on new diseases is scarce. The XDT framework employs vision encoder embeddings and kernel transformation to differentiate between diseased and non-diseased states, making it especially beneficial in resource-poor settings. However, it currently only supports binary classification, identifying disease presence or absence without distinguishing between multiple diseases. Despite this, the XDT-CXR framework surpasses other zero-shot learning baselines, highlighting its potential as a valuable diagnostic tool.

Acknowledgments

We would like to acknowledge the discussions with Dr. Jennifer Kate van Heerden from the Nuffield Department of Surgical Sciences, University of Oxford, Oxford. Her insights and expertise significantly contributed to the development and refinement of this work.

References

Jameel Abdul Samadh, Mohammad Hanan Gani, Noor Hussein, Muhammad Uzair Khattak, Muhammad Muzammal Naseer, Fahad Shahbaz Khan, and Salman H Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. *Advances in Neural Information Processing Systems*, 36, 2024.

A.S. Al-Waisy, S. Al-Fahdawi, and M.A. Mohammed. Covid-chexnet: hybrid deep learning framework for identifying covid-19 virus in chest x-rays images. *Soft Computing*, 27: 2657–2672, 2023. doi: https://doi.org/10.1007/s00500-020-05424-3.

Bang An, Sicheng Zhu, Michael-Andrei Panaitescu-Liess, Chaithanya Kumar Mummadi, and Furong Huang. More context, less distraction: Improving zero-shot inference of clip by inferring and describing spurious features. In *Workshop on Efficient Systems for Foundation Models@ ICML2023*, 2023.

Thomas M Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers*, (3): 326–334, 1965.

Xianfan Gu, Yu Zhang, Wen Zeng, Sihua Zhong, Haining Wang, Dong Liang, Zhenlin Li, and Zhanli Hu. Cross-modality image translation: Ct image synthesis of mr brain images using multi generative network with perceptual supervision. *Computer Methods and Programs in Biomedicine*, 237:107571, 2023.

Namig J Guliyev and Vugar E Ismailov. On the approximation by single hidden layer feedforward neural networks with fixed weights. *Neural Networks*, 98:296–304, 2018.

Nasir Hayat, Hazem Lashen, and Farah E Shamout. Multi-label generalized zero shot learning for the classification of disease in chest radiographs. In *Machine learning for healthcare conference*, pages 461–477. PMLR, 2021a.

Nasir Hayat, Hazem Lashen, and Farah E. Shamout. Multi-label generalized zero shot learning for the classification of disease in chest radiographs, 2021b.

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

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

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.

Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5): 1122–1131, 2018.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023a.

Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15190–15200, 2023b.

Muhammad Uzair Khattak, Muhammad Ferjad Naeem, Muzammal Naseer, Luc Van Gool, and Federico Tombari. Learning to prompt with text only supervision for vision-language models. *arXiv preprint arXiv:2401.02418*, 2024.

Rohit Kundu, Ritacheta Das, Zong Woo Geem, Gi-Tae Han, and Ram Sarkar. Pneumonia detection in chest x-ray images using an ensemble of deep learning models. *PloS one*, 16 (9):e0256630, 2021.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553): 436–444, 2015.

Muhammad Jehanzeb Mirza, Leonid Karlinsky, Wei Lin, Horst Possegger, Mateusz Kozinski, Rogerio Feris, and Horst Bischof. Lafter: Label-free tuning of zero-shot classifier using language and unlabeled image collections. *Advances in Neural Information Processing Systems*, 36, 2024.

Majid Nour, Zafer Cömert, and Kemal Polat. A novel medical diagnosis model for covid-19 infection detection based on deep features and bayesian optimization. *Applied Soft Computing*, 97:106580, 2020. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2020.106580. URL https://www.sciencedirect.com/science/article/pii/S1568494620305184.

Hasan Selçuk Özger, PINAR AYSERT YILDIZ, Ümmügülsüm Gaygisiz, Asİye Uğraş Dikmen, Zehra Demirbaş Gülmez, Mehmet Yildiz, Esin Şenol, Kenan Hizel, Özlem Güzel Tunçcan, Kayhan Çağlar, et al. The factors predicting pneumonia in covid-19 patients: preliminary results from auniversity hospital in turkey. *Turkish journal of medical sciences*, 50(8):1810–1816, 2020.

Angshuman Paul, Thomas C. Shen, Sungwon Lee, Niranjan Balachandar, Yifan Peng, Zhiyong Lu, and Ronald M. Summers. Generalized zero-shot chest x-ray diagnosis through trait-guided multi-view semantic embedding with self-training. *IEEE Transactions on Medical Imaging*, 40(10):2642–2655, 2021. doi: 10.1109/TMI.2021.3054817.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021a. URL https://proceedings.mlr.press/v139/radford21a.html.

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

Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam, Somaya Al Maadeed, Susu M Zughaier,

Muhammad Salman Khan, et al. Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Computers in biology and medicine*, 132: 104319, 2021.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.

Richard Souvenir, Qi Zhang, and Robert Pless. Image manifold interpolation using free-form deformations. In *Proc. IEEE International Conference on Image Processing*, pages 1437–1440, 2006.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Qi Zhang, Richard Souvenir, and Robert Pless. Segmentation informed by manifold learning. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, 2005.

Qi Zhang, Richard Souvenir, and Robert Pless. On manifold structure of cardiac mri data: Application to segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1092–1098, 2006.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.

Appendix

Appendix A. Choice of network architecture

Table 5 below shows ablation experiments using a multi-layer perceptron (MLP) block and a linear layer instead of a transformer block, resulting in weaker performance on medical tasks. Experiments with a ResNet model without MedCLIP highlight the importance of using a MedCLIP image encoder with a transformer for better cross-disease transferability.

Table 5: Performance metrics for various models across different training and testing sets.

Model	Train	S	C	G	M
MedCLIP + Linear	S	42.86	26.25	62.5	48.28
	C	42.86	26.25	62.5	48.28
	G	42.86	26.25	62.5	48.28
	M	42.86	26.25	62.5	48.28
MedCLIP + MLP	S	57.14	73.75	37.5	51.72
	C	57.14	73.75	37.5	51.72
	G	57.14	73.75	37.5	51.72
	M	57.14	73.75	37.5	51.72
ResNet	S	89.47	63.85	43.43	62.07
	C	63.16	93.93	76.28	68.97
	G	48.87	49.67	37.50	51.72
	M	60.90	48.30	57.05	68.97
Statistical Best	x	54.17	73.75	62.5	51.72

Appendix B. Choice of Visual Encoder

Table 6 provides performance metrics for different models and architectures, comparing the TPT and LaFter models using the two different visual encoders: CLIP and MedCLIP. The metrics are reported for three categories: G, S, and M. For the TPT model, the MedCLIP encoder generally outperforms the CLIP encoder across all categories, with notable improvements in categories S and M. Specifically, MedCLIP boosts the performance in category S (54.08%) and M (57.97%) compared to CLIP's lower scores (50.76% and 42.03%, respectively). The choice of encoder impacts model performance variably across different categories. Since the overall accuracy of MedCLIP was more than CLIP, we selected MedCLIP as the visual encoder for all our experiments. performs better comparatively.

Appendix C. Adapting Baseline Methods to CXR Datasets

Table 8 presents the performance of the LaFter model on the test sets of different datasets: Shenzhen-TB (S), Covid (C), Guangzhou-PN (G), and Montgomery-TB (M). For each training dataset (denoted as Train), the table shows the model's accuracy across the different evaluation datasets. For instance, when LaFter is trained on Shenzhen-TB and evaluated on the same (S), it achieves an accuracy of 57.14%. Conversely, when trained on Covid (C) and

Table 6: Performance metrics for various models and architectures

Model	Arch	G	S	M
TPT	CLIP	72.97	50.76	42.03
	MedCLIP	72.98	54.08	57.97
LaFter	CLIP	62.5	57.14	51.72
	MedCLIP	78.53	42.86	48.28

Table 7: Overall accuracy comparison between CLIP and MedCLIP shows that MedCLIP performs better comparatively.

CLIP	MedCLIP
337.12	354.7

evaluated on Guangzhou-PN (G), it achieves a higher accuracy of 76.28%. The diagonal values reported in Table 2, which are not shown here, likely reflect the model's performance when trained and evaluated on the same dataset, providing a baseline for comparison. This table helps illustrate how well LaFter adapts across various CXR datasets, highlighting its strengths and weaknesses in different training and evaluation scenarios.

Table 8: Performance was evaluated on the test set for each dataset in Section 4.1, with Table 2 reporting LaFter's diagonal values.

Model	Train	S	C	G	M
LaFter	S	57.14	73.75	37.5	51.72
	C	42.86	26.25	76.28	42.86
	G	42.86	26.25	62.5	48.28
	M	57.14	73.75	37.5	51.72

Appendix D. Choice of Loss Function

Table 9 ranks the accuracy of various models trained on different datasets and evaluated with different loss functions. For each train-eval pair (where the training dataset and evaluation dataset are denoted as S: Shenzhen-TB, C: Covid, G: Guangzhou-PN, M: Montgomery-TB), the table lists the ranking of accuracy across four loss functions: $\mathcal{L}_{\text{EC}}$, $\mathcal{L}_{\text{CE}}$, $\mathcal{L}_{\text{LC}}$ with $\lambda = 0$, and $\mathcal{L}_{\text{LC}}$ with $\lambda = 0.001$. Rankings are assigned from 0 (best) to 4 (worst). For example, when trained on Shenzhen-TB and evaluated on Covid, $\mathcal{L}_{\text{EC}}$ is ranked 1st, whereas $\mathcal{L}_{\text{CE}}$ is ranked 4th in the same scenario. The table provides insights into how different loss functions perform across various dataset combinations, highlighting which configurations yield the highest accuracy.

Then, we have calculated the mean average rank (MAR) for each loss function, and observe that logarithmic contrastive (LC) loss has the lowest MAR, and therefore we choose it for our framework.

Table 9: Ranked accuracy for each train-eval pair across all loss functions.

Trained on	Loss	S	C	G	M
S	$\mathcal{L}_{\text{EC}}$	0	1	4	1
C		2	0	3	2
G		2	2	0	2.5
M		1	3	1	0
S	$\mathcal{L}_{\text{CE}}$	0	4	3	4
C		4	0	4	4
G		3.5	4	0	4
M		4	4	4	0
S	$\mathcal{L}_{\text{LC}}, \lambda = 0$	0	2	2	2.5
C		1	0	2	1
G		3.5	3	0	2.5
M		2.5	2	2.5	0
S	$\mathcal{L}_{\text{LC}}, \lambda = 0.001$	0	3	1	2.5
C		3	0	1	3
G		1	1	0	1
M		2.5	1	2.5	0

Table 10: Mean average rank (MAR) for each loss function.

Loss	MAR
$\mathcal{L}_{\text{EC}}$	2.04
$\mathcal{L}_{\text{CE}}$	3.88
$\mathcal{L}_{\text{LC}}, \lambda = 0$	2.21
$\mathcal{L}_{\text{LC}}, \lambda = 0.001$	1.88