Supplementary Material for Learning Anatomy-Disease Entangled Representation

Fatemeh Haghighi[1] Michael B. Gotway[2]
Jianming Liang[1]
[1]Arizona State University, USA [2]Mayo Clinic, USA
{fhaghigh,jianming.liang}@asu.edu Gotway.Michael@mayo.edu

A. Datasets and downstream tasks

Tab. 1 in the main paper and Tab. 4 in the Appendix summarize the datasets used for training LeADER and the downstream tasks on which LeADER has been evaluated, respectively. Below, we provide the details of each dataset as well as the downstream tasks associated with each dataset.

PadChest [1]**:** The PadChest dataset is a large-scale, high-resolution chest X-ray dataset that includes over 160,000 chest X-ray images from 67,000 patients, collected between 2009 and 2017 at Hospital San Juan de Dios in Spain. The dataset features six different positional views and provides additional information on image acquisition and patient demographics. The reports associated with the images are labeled with 174 radiographic findings, 19 differential diagnoses, and 104 anatomical locations, which are organized into a hierarchical taxonomy and mapped to standard Unified Medical Language System (UMLS) terminology. Of these reports, 27% were manually annotated by trained physicians, while the remaining annotations were generated using a supervised method based on a recurrent neural network with attention mechanisms. This dataset is notable for being the first to include radiographic reports in Spanish.

ChestX-ray14 [13]**:** The NIH ChestX-ray14 dataset is a hospital-scale collection comprising 112,120 frontal view X-ray images from 32,717 unique patients. Labels for 14 common thoracic pathologies are extracted from the chest X-ray radiological reports using natural language processing techniques, with each image potentially having multiple labels. The dataset provides an official patient-wise split for training (86,000 images) and test (25,000 images) sets. In our study, we use the ChestX-ray14 dataset both as a pretraining source and as a target dataset. We follow the official data split and report the mean AUC score over 14 diseases for the multi-label chest X-ray classification task.

CheXpert [4]**:** The CheXpert dataset is a large-scale collection of 223,414 multi-view chest radiographs from 65,240 patients. The training images were annotated by an automated labeler to detect the presence of 14 thoracic diseases in radiology reports, accounting for uncertainties with an uncertainty label. The validation set comprises 234 images from 200 patients, manually annotated by board-certified radiologists for 5 selected diseases. For pretraining, we just use the training samples based on the official split. We fol-

Downstream Tasks			
Dataset	Downstream task	#Train/#Test/Data split	Metric
ChestX-ray14	14 thoracic diseases classification	86K/25K/Official	AUC
CheXpert	14 thoracic diseases classification	223K/234/Official	AUC
Shenzhen CXR	Tuberculosis classification	529/133/Random	AUC
VinDr-CXR	14 thoracic diseases classification	15K/3K/Official	AUC
SIIM-ACR	Pneumothorax segmentation	8K/2K/Random	Dice
RSNA Pneumonia	Pneumonia detection	-/2709/Random	Precision
COVIDx	COVID-19 classification	29633/400/Official	Accuracy
JSRT	Lung nodule classification	197/50/Random	AUC
ChestX-Det	13 thoracic diseases classification	3K/553/Official	AUC
ChestX-Det	13 thoracic diseases segmentation	3K/553/Official	IoU

Table 4. Summary of downstream tasks on which LeADER has been evaluated.

low the official data split and report the mean AUC score over 5 test diseases for the multi-label chest X-ray classification task.

NIH Shenzhen CXR [5]**:** The NIH Shenzhen CXR dataset includes 662 frontal-view chest X-rays, with 326 normal cases and 336 showing Tuberculosis (TB) manifestations. We randomly split the dataset into a training set (80%) and a test set (20%). The AUC score for the Tuberculosis detection task is reported.

RSNA Pneumonia [12]**:** The RSNA Pneumonia dataset, derived from the RSNA Pneumonia Detection Challenge 2018, consists of 30,000 frontal view chest radiographs from the public National Institutes of Health (NIH) CXR8 dataset. This includes 16,248 posteroanterior views and 13,752 anteroposterior views, with a separate test set of 4,527 images. The annotations for this dataset were provided by 18 radiologists from 16 different institutions, including 12 chest radiologists from the STR Specialty, with an average of 10.6 years of experience (ranging from 3 to 35 years). The annotations include bounding boxes and are categorized as follows: 0 - Unknown, 1 - Pneumonia present, 2 - Pneumonia absent.

MIMIC-CXR [6]**:** MIMIC-CXR dataset is a large, publicly available dataset that includes chest radiographs along with their corresponding radiological reports. The dataset comprises 377,110 images from 227,835 radiographic studies and offers image-level labels for 13 thoracic diseases, which are derived from the radiology reports through the use of two open-source labeling tools, NegBio and CheXpert. The images are split into 368,000 for training, 2,991 for validation, and 5,159 for testing in the official data split. For our study, we leverage the MIMIC-CXR dataset for pretraining the source model using the official training data.

VinDR-CXR [10]**:** The VinDR-CXR dataset consists of 36,096 posterior-anterior chest X-rays with image-level labels assigned by expert radiologists for six conditions: lung tumor, pneumonia, tuberculosis, other diseases, COPD, and no finding. Additionally, the dataset includes bounding box labels for 14 conditions such as Aortic enlargement, Atelectasis, Calcification, Cardiomegaly, Consolidation, ILD, Infiltration, Lung Opacity, Nodule/Mass, Other lesion, Pleural effusion, Pleural thickening, Pneumothorax, and Pulmonary fibrosis. For pretraining, we use only the training samples based on the official split. For fine-tuning, we adhere to the official data split and report the mean AUC score over the six diseases for the image-level classification task and the AUC over 14 diseases for the lesion-level classification task.

ChestX-Det [7]**:** The ChestX-Det dataset comprises 3,578 chest X-ray images, each annotated with pixel-level segmentation masks for 13 common thoracic conditions: atelectasis, calcification, cardiomegaly, consolidation, diffuse nodule, effusion, emphysema, fibrosis, fracture, mass, nodule, pleural thickening, and pneumothorax. The official dataset split includes 3,025 images for training and 553 images for testing. For the 13 thoracic segmentation task, we report the Intersection over Union (IoU) score, while for the lesion-level classification task, we report the AUC score across the 13 diseases.

Node21 [11]**:** The Node21 public CXR training dataset is a collection of frontal chest X-ray images specifically designed for training and evaluating systems in both detection and generation tasks. The dataset includes a total of 4,882 frontal chest radiographs, sourced from four public datasets: JSRT, PadChest, ChestX-ray14, and Open-I, all of which permit remixing and redistribution of the images. Out of these images, 1,134 are annotated with bounding boxes around 1,476 nodules, while the remaining 3,748 images are negative samples, meaning they do not contain any nodules. The annotations for the dataset were provided by experienced chest radiologists.

TBX-11K [8]**:** The TBX-11K dataset consists of 11,200 chest radiographs, each meticulously annotated with bounding boxes to identify tuberculosis (TB) areas. It categorizes images into five distinct classes: Healthy, Sick but Non-TB, Active TB, Latent TB, and Uncertain TB. Out of the 11,200 X-rays, there are 5,000 healthy cases and 5,000 sick but non-TB cases. Additionally, there are 1,200 cases showing manifestations of TB, with each chest radiograph representing a unique individual. Within these TB cases, there are 924 instances of active TB, 212 cases of latent TB, 54 cases where both active and latent TB are present simultaneously, and 10 cases classified as uncertain, where the TB type cannot be identified using current medical standards. The official split of the TBX11K dataset is as follows: the training set comprises 6,600 images, the validation set includes 1,800 images, and the testing set contains 2,800 images.

SIIM-ACR [15]**:** The SIIM-ACR dataset consists of 10,000 chest X-ray images and segmentation masks for Pneumothorax disease, provided by the Society for Imaging Informatics in Medicine (SIIM) and the American College of Radiology. We randomly divided the dataset into training (80%) and testing (20%) sets, and evaluated the segmentation performance using the Dice coefficient score.

B. Implementation details

B.1. Pretraining settings

As summarized in Tab. 5, we trained LeADER using 900K samples collected from the training sets of 10 public datasets [1, 4–8, 10–13], with the Swin transformer base

Training LeADER Settings	
Backbone	S_θ: Swin-B transformer
Heads	h_{θ_D} and h_{θ_A}: 2-layers MLP with hidden-dim 2048
Input Resolution	224×224
Initialization	S_θ: officially released ImageNet weights; h_{θ_D} and h_{θ_A}: random initialization
Loss	$\mathcal{L}_D$ and $\mathcal{L}_A$: Mean Squared Error (MSE)
Batch size	128 distributed across 4 Nvidia V100 GPUs
Augmentation	Random affine transformation, horizontal flip, and color jitter
Optimization	AdamW optimizer with learning rate $2e-4$ cosine annealing learning rate scheduler, warm-up epochs 10
Training	100 epochs with $\mathcal{L}_D$ and 250 epochs with $\mathcal{L}_D + \mathcal{L}_A$

Table 5. Summary of pretraining settings

(Swin-B) [9] as the backbone of the student S_θ. The two-layer MLP heads for h_{θ_D} and h_{θ_A} have a hidden dimension of 2048 and output dimensions of $K = 1376$ and $K = 2048$ for the disease and anatomy heads, respectively. For the disease expert T_{ξ_D}, we employ Google CXR-FM, known for its proficiency in generating disease-related features. For the anatomy expert T_{ξ_A}, we employ Adam [2], which excels in generating anatomy-related features. It should be noted that other suitable disease/anatomy expert models can also be integrated into our framework without constraint. During training, we optimize S_θ, h_{θ_D}, and h_{θ_A} using the AdamW optimizer with a base learning rate of $2e-4$, 10 warm-up epochs, and cosine annealing learning rate scheduler, while T_{ξ_D} and T_{ξ_A} are kept frozen. Mean Squared Error (MSE) is used as $\mathcal{L}_D$ and $\mathcal{L}_A$. Random affine transformation, horizontal flip, and color jitter are used as data augmentation. LeADER's backbone (Swin-B model) is initialized from the officially released ImageNet weights, while both heads, h_{θ_D} and h_{θ_A}, are randomly initialized. LeADER is trained using a batch size of 128 with 4 Nvidia V100 GPUs. For 100 epochs, LeADER is trained using the PadChest [1], ChestX-ray14 [13], CheXpert [4], NIH Shenzhen CXR [5], RSNA Pneumonia [12], MIMIC-CXR [6], and VinDR-CXR [10] datasets with $\mathcal{L}_D$ to develop an initial disease discrimination ability. Following this, for 250 epochs, LeADER is further enhanced by joint optimization of $\mathcal{L}_D$ and $\mathcal{L}_A$ using both images and patches from the VinDR-CXR, ChestX-Det, NODE21, and TBX11K [8] datasets, enabling the model to capture rich entangled and disentangled representations at both image and patch levels.

B.2. Fine-tuning settings

As summarized in Tab. 6, we utilize the backbone of the pretrained student of LeADER (i.e., S_θ) for full transfer evaluation. In transfer learning to classification tasks, we take LeADER's pretrained backbone and append a fully connected layer to generate the task-specific classification outputs; with batch size 32, we use SGD optimizer with learning rate $1e-2$, decrease the learning rate with a cosine scheduler, and use standard data augmentation, encompassing random rotation, crop, and horizontal flip. In transfer learning to segmentation tasks, we employ a UperNet architecture [14], initializing the encoder weights with LeADER's pretrained backbone; with batch size 32, we use

Transfer Learning Settings	
Input resolution	224×224
Augmentation	cls[1]: random rotation, crop, and horizontal flip seg[1]: random gamma, elastic, brightness contrast, optical & grid distortion
Optimization	batch size: 32; optimizer: SGD; lr: 1e-2/1e-3 for cls/seg; learning rate decay scheduler: cosine for cls/seg;
Architecture	cls: Swin-B followed by a task-specific classification head seg: UperNet with Swin-B encoder
Transferred model	Student S_θ
Statistical Analysis	Independent two-sample t-test at $p = 0.05$ level

[1] "cls" and "seg" denote classification and segmentation tasks, respectively.

Table 6. Summary of transfer learning settings

SGD optimizer with learning rate $1e-3$, decayed by a cosine schedule, and use standard data augmentation, encompassing random gamma, elastic, brightness contrast, optical and grid distortion. Following the standard evaluation protocol [3], we perform end-to-end fine-tuning for all parameters of the target models across all downstream tasks. We strive to optimize each downstream task with the most effective hyperparameters. Moreover, we employ early-stopping using 10% of the training data as the validation set. We use input size 224^2 for all downstream tasks. Classification and segmentation performances are measured by the AUC (area under the ROC curve), and mean Dice coefficient metrics and IoU (Intersection over Union) metrics, respectively. We run each downstream model at least five times and report statistical analysis using an independent two-sample t-test.

C. Computational efficiency

LeADER uses a shared student backbone for both anatomy and disease branches, with lightweight learning heads for each branch. Additionally, the teacher models remain frozen during training. As a result, LeADER does not impose additional computational demands compared to the baselines and is even less resource-intensive than multi-task methods like DiRA and PCRL, which require substantial negative pairs comparisons for their contrastive learning objectives.

D. Ethical considerations

AI in medicine is still in an early stage. Hence, any commercial deployment of the models presented in our research study should not proceed without sufficient evaluations in real-world clinical settings.

E. Acknowledgements

This research was partially supported by ASU and Mayo Clinic through a Seed Grant and an Innovation Grant, as well as by the NIH under Award Number R01HL128785. The authors are solely responsible for the content, which does not necessarily reflect the official views of the NIH. This work also utilized GPUs provided by ASU Research Computing, Bridges-2 at the Pittsburgh Supercomputing Center (allocated under BCS190015), and Anvil at Purdue University (allocated under MED220025). These resources are supported by the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, funded by the National Science Foundation under grants #2138259, #2138286, #2138307, #2137603, and #2138296. The content of this paper is covered by patents pending.

References

[1] Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797, 2020. 1, 2, 3

[2] Mohammad Reza Hosseinzadeh Taher, Michael B. Gotway, and Jianming Liang. Towards foundation models learned from anatomy in medical imaging via self-supervision. In *Domain Adaptation and Representation Transfer*, pages 94–104, Cham, 2024. Springer Nature Switzerland. 3

[3] Mohammad Reza Hosseinzadeh Taher, Fatemeh Haghighi, Ruibin Feng, Michael B. Gotway, and Jianming Liang. A systematic benchmarking analysis of transfer learning for medical image analysis. In *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*, pages 3–13, Cham, 2021. Springer International Publishing. 4

[4] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, and et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *arXiv:1901.07031*, 2019. 1, 2, 3

[5] 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), 2014. 2, 3

[6] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6:317, 2019. 2, 3

[7] Jie Lian, Jingyu Liu, Shu Zhang, Kai Gao, Xiaoqing Liu, Dingwen Zhang, and Yizhou Yu. A structure-aware relation network for thoracic diseases detection and segmentation. *IEEE Transactions on Medical Imaging*, 40(8):2042–2052, 2021. 2

[8] 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. 2, 3

[9] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, October 2021. 3

[10] Ha Q. Nguyen and et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations, 2020. 2, 3

[11] Ecem Sogancioglu, Bram van Ginneken, Finn Behrendt, Marcel Bengs, Alexander Schlaefer, and et al. Nodule detection and generation on chest x-rays: Node21 challenge, 2024. 2

[12] Anouk Stein, Carol Wu, Chris Carr, and et al. Rsna pneumonia detection challenge, 2018. 2, 3

[13] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks

on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, pages 2097–2106, 2017. 1, 2, 3

[14] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 3

[15] Anna Zawacki, Carol Wu, George Shih, Julia Elliott, Mikhail Fomitchev, Mohannad Hussain, ParasLakhani, Phil Culliton, and Shunxing Bao. Siim-acr pneumothorax segmentation, 2019. 2