Leveraging Pathology Co-occurrence for Test-Time Adaptation in Chest X-Ray Diagnosis

Woojin Jeong[1†], Yujin Choi[2,3†], Dongbin Kim[1], Soyeon Park[1], and Jaewook Lee[1(✉)]

[1] Seoul National University, 1, Gwanak-ro, Seoul, 08826, Republic of Korea
[2] Nanyang Technological University, 50 Nanyang Avenue, 639798, Singapore
[3] UNIST, 50, UNIST-gil, Ulsan, 44919, Republic of Korea
jaewook@snu.ac.kr

Abstract. Medical imaging models often degrade when deployed at new clinical sites due to differences in imaging equipment, protocols, and patient populations. Test-time adaptation (TTA) addresses this by updating a pretrained model using only unlabeled target data, without access to source data. However, existing TTA methods were designed for single-label classification on natural image benchmarks, minimizing entropy uniformly across all samples without considering label dependencies. This overlooks a key property of multi-label medical imaging: pathologies do not occur independently but exhibit structured co-occurrence patterns. In this work, we propose **Co**-occurrence **W**eighted **A**daptation (CoWA), which leverages disease co-occurrence patterns as a reliability signal for adaptation. CoWA estimates label co-occurrence structure from model predictions and downweights samples that deviate from expected patterns, enabling adaptation to rely more on consistent predictions while reducing the impact of noisy ones. We evaluate CoWA on chest X-ray benchmarks under domain shifts and demonstrate consistent improvements over established baselines. The code is available at https://github.com/woojin716/CoWA.

Keywords: Chest X-ray · Domain Shift · Test-time Adaptation

1 Introduction

Deep learning models for chest X-ray classification have demonstrated strong performance on benchmark datasets [6, 17]. In clinical practice, however, models are often deployed across institutions where imaging equipment, acquisition protocols, and patient demographics differ from those in the training environment [10, 13, 20]. The resulting domain shift can severely degrade diagnostic accuracy, limiting the real-world adoption of automated systems [4]. Addressing this performance gap under domain shift remains a key challenge for reliable clinical AI deployment.

† Equal contribution.

To address domain shift, prior work has explored domain adaptation [3] and domain generalization [21], both of which require access to source data or control over the training process. In practice, pretrained models are frequently distributed to new sites where source data is unavailable due to privacy regulations and institutional policies. Test-time adaptation (TTA) provides a more realistic alternative by updating a pretrained model at inference using only unlabeled target data, without any source data or target annotations [15].

While TTA has been widely studied for natural image classification, its direct application to multi-label medical imaging remains problematic. Standard TTA minimizes prediction entropy per class and weights all samples equally, implicitly assuming label independence. However, clinical pathologies follow structured co-occurrence patterns rather than occurring in isolation [17, 18]. Fig. 1 shows that these dependency structures vary across datasets, reflecting differences in clinical setting. Predictions that violate plausible label relationships may therefore be unreliable despite being individually confident. Yet existing entropy-based TTA methods adapt using all predictions indiscriminately, amplifying inconsistent patterns and introducing noisy gradients under domain shift.

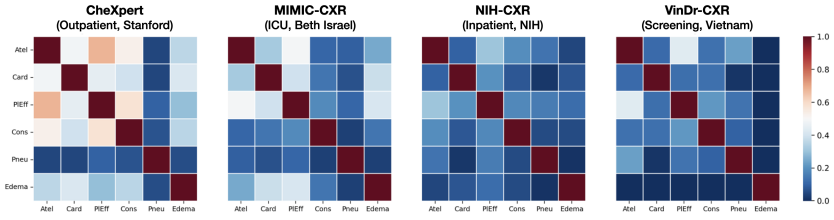

Fig. 1. Co-occurrence patterns vary across chest X-ray domains.

In this work, we propose **CoWA** (**Co**-occurrence **W**eighted **A**daptation) to bridge this gap. CoWA estimates the pathology co-occurrence matrix from incoming test predictions, capturing the evolving co-occurrence patterns of the target domain without accessing source data. This matrix serves as a reliability signal that measures how well each sample's predicted label configuration aligns with the emerging target structure. During adaptation, CoWA assigns higher weights to structurally consistent samples and downweights predictions that violate plausible label relationships. By steering updates toward coherent label patterns, CoWA reduces noisy gradients and enables TTA to adapt while respecting structured dependencies in multi-label clinical data, without requiring additional supervision.

Our main contributions can be summarized as follows:

1. We propose CoWA, a source-free TTA method that estimates co-occurrence patterns from model predictions and uses them as a per-sample reliability signal for multi-label medical image classification.

2. We show that standard TTA methods, originally designed for single-label benchmarks, can be ineffective or even harmful when applied to multi-label clinical tasks, highlighting the need for label-structure aware adaptation.
3. We validate CoWA on multiple domain shift scenarios using public chest X-ray datasets, showing consistent gains over established baselines.

2 Related Work

Test-time adaptation (TTA) has emerged as an effective strategy for addressing distribution shifts, as it adapts models to the target distribution without access to source data, unlike traditional domain adaptation. Existing TTA methods update model parameters via prediction-based objectives. TENT [15] minimizes prediction entropy by updating BN affine parameters, while EATA [12] extends this by filtering unreliable samples and applying Fisher regularization to mitigate catastrophic forgetting. CoTTA [16] and RoTTA [19] further address continual adaptation under temporally correlated streams via augmentation-based pseudo-labels and robust BN estimation, respectively. Beyond gradient-based optimization, AdaBN [8] replaces source BN statistics with those estimated from target samples, serving as a simple yet effective domain correction baseline. While these methods have shown success on single-label natural image benchmarks such as ImageNet-C [5], they typically treat all samples equally during adaptation and assume label independence. In multi-label medical imaging, however, chest pathologies exhibit structured co-occurrence patterns [9, 1], suggesting that label dependencies can provide valuable guidance for adaptation.

3 Proposed Method

Given a source-pretrained multi-label classifier f_θ and an unlabeled target dataset $\mathcal{D}_t = \{x_i\}_{i=1}^N$, CoWA adapts f_θ to the target distribution by incorporating target domain label co-occurrence structure into entropy minimization. Instead of uniformly minimizing entropy over all predictions, CoWA estimates the target co-occurrence structure from model outputs and assigns a weight to each sample based on its structural consistency. These weights are then used in entropy minimization during adaptation.

CoWA consists of three components: (1) estimation of the target domain co-occurrence matrix from model predictions, (2) computation of a sample-wise consistency score, and (3) weighted entropy minimization to update model parameters. The overall framework is illustrated in Fig. 2.

Co-occurrence Matrix Estimation. For an input x_i, the model outputs $\hat{y}_i = f_\theta(x_i) \in [0,1]^C$, where C denotes the number of pathologies. We treat these predictions as soft pseudo-labels and obtain $\tilde{y}_i \in \{0,1\}^C$ via fixed threshold binarization to suppress low confidence activations and reduce noise in dependency estimation. Based on these binary predictions, we compute the pairwise

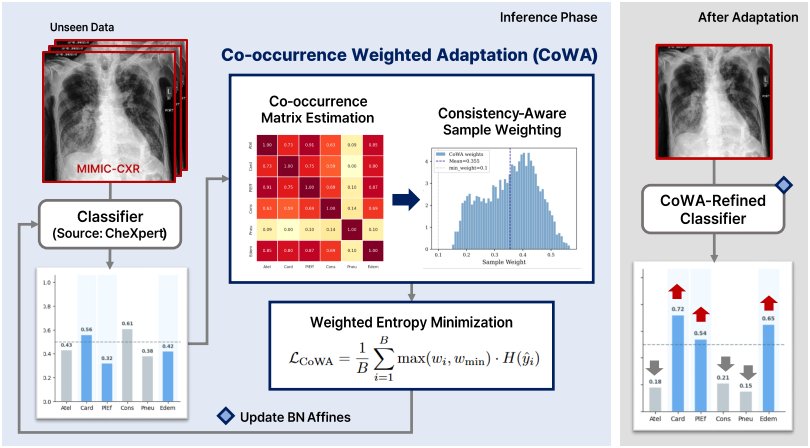

Fig. 2. Overview of CoWA. CoWA estimates a pathology co-occurrence matrix from target predictions, computes sample-wise consistency scores, and uses them to weight samples for test-time domain adaptation.

co-occurrence matrix:

$$\mathbf{S} = \sum_{i=1}^{n} \tilde{y}_i \tilde{y}_i^{\top}, \tag{1}$$

where n is the number of processed samples and $\mathbf{S}_{jk}$ counts how often pathologies j and k are jointly predicted as positive. We then compute the empirical joint probabilities $\mathbf{P} = \mathbf{S}/n$ and define the normalized co-occurrence matrix as

$$\mathbf{M}_{jk} = \frac{\mathbf{P}_{jk}}{\sqrt{\mathbf{P}_{jj}\mathbf{P}_{kk}} + \varepsilon}, \quad \mathbf{M}_{jj} = 1. \tag{2}$$

Here, $\mathbf{S}$ is accumulated across all batches processed so far rather than recomputed per batch, so that $\mathbf{M}$ is progressively refined as adaptation proceeds and the influence of noisy early-stage predictions is dampened.

Consistency-aware Sample Weighting. Using the estimated co-occurrence matrix $\mathbf{M}$, we assign a consistency score to each target prediction. Here, the consistency of a prediction measures how well its implied label relationships match the estimated target co-occurrence structure. For a sample with predicted probabilities $\hat{y}_i$, we construct a local co-occurrence pattern $\mathbf{m}_i = \hat{y}_i \hat{y}_i^{\top}$ and compare it with $\mathbf{M}$ using the Frobenius norm, converting the deviation into a reliability weight:

$$w_i = \exp\left(-\frac{\|\mathbf{m}_i - \mathbf{M}\|_F^2}{\tau}\right), \tag{3}$$

where τ is a temperature parameter controlling the weighting. Predictions with higher consistency get weights close to 1, while lower-consistency predictions are

assigned smaller weights, effectively reducing the influence of samples that are less aligned with the target co-occurrence structure.

Weighted Entropy Minimization. Existing entropy-based TTA methods [15, 12] adapt models by minimizing the uniformly averaged prediction entropy, treating all test samples equally. This implicitly assumes that every confident prediction is equally reliable, regardless of its structural consistency with the target domain. In contrast, CoWA reweights each sample based on its consistency score and adjusts its contribution to the adaptation objective:

$$\mathcal{L}_{\text{CoWA}} = \frac{1}{B}\sum_{i=1}^{B} \max(w_i, w_{\min}) \cdot H(\hat{y}_i), \tag{4}$$

where $H(\cdot)$ denotes the entropy of the model prediction, B is the batch size, and $w_{\min}$ prevents the weights from collapsing to zero for numerical stability, acting as an early-stage safeguard before $\mathbf{M}$ stabilizes. Samples whose predictions align well with the estimated target co-occurrence structure receive larger weights and thus exert stronger influence on adaptation, whereas structurally inconsistent samples are suppressed. Following prior work [15], we adapt only the BN affine parameters of f_θ.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on four public chest X-ray datasets: MIMIC-CXR [7], CheXpert [6], VinDr-CXR [11], and NIH Chest X-ray [17]. We construct six source–target pairs to cover diverse domain shift scenarios. For each pair, the unseen test split of the target domain is used for evaluation. Results are reported on six shared pathologies: Atelectasis (At.), Cardiomegaly (Ca.), Pleural Effusion (Ef.), Consolidation (Co.), Pneumothorax (Pt.), and Edema (Ed.).
Pretrained Classifiers. We use DenseNet-121 backbones from TorchXRayVision [2], pretrained on NIH and CheXpert. Each pretrained model is treated as a distinct source model when evaluating cross-domain shifts.
Baselines. We evaluate CoWA against *No Adapt.* (source model without adaptation), AdaBN [8], and four TTA methods: TENT [15], CoTTA [16], EATA [12], and RoTTA [19].
Implementation Details. All methods use a batch size of 64, with baseline hyperparameters following their original papers. For CoWA, the learning rate is selected from [1e-4, 1e-2], with $\tau \in \{0.01, 0.05, 0.1, 0.5\}$, and the co-occurrence threshold $\in \{0.4, 0.5, 0.6\}$. We fix $w_{\min} = 0.01$ in all settings. All experiments are conducted on four NVIDIA GeForce RTX 3090 GPUs.

4.2 Main Results

Table 1 reports the AUROC across six domain shift scenarios. CoWA achieves the best mean performance in most settings and remains competitive in the

Table 1. AUROC under multiple domain-shift settings across source domains. **Bold** indicates the best result and underline the second best. Values are scaled by 100 for readability.

Method	Source: CheXpert							Source: NIH						
	Pathology						Mean	Pathology						Mean
	At.	Ca.	Ef.	Co.	Pt.	Ed.		At.	Ca.	Ef.	Co.	Pt.	Ed.	
	Target: MIMIC													
No Adapt.	61.8	69.5	73.8	58.7	63.0	73.6	66.7	62.6	66.3	73.9	57.0	**57.5**	67.8	64.2
AdaBN	63.4	68.5	**76.0**	**62.2**	65.9	72.8	68.1	63.5	68.9	73.9	57.1	48.3	69.7	63.6
TENT	**64.8**	68.0	75.5	59.4	64.5	73.7	67.6	62.8	65.2	71.2	54.8	49.1	62.8	61.0
CoTTA	61.9	69.5	73.8	58.6	63.0	73.6	66.7	62.6	66.4	73.9	57.0	57.3	67.9	64.2
EATA	63.3	68.5	75.6	61.5	65.8	72.6	67.9	64.6	**69.6**	**74.8**	**57.8**	49.5	72.0	64.7
RoTTA	61.8	**69.6**	73.8	58.6	62.9	73.6	66.7	62.5	66.2	73.8	56.9	**57.5**	67.7	64.1
CoWA	64.2	68.6	**76.0**	61.3	**66.3**	**73.8**	**68.4**	**65.1**	68.9	73.3	**57.8**	55.7	**72.1**	**65.5**
	Target: VinDr													
No Adapt.	77.7	**85.3**	82.6	81.0	75.7	–	80.5	74.9	82.4	85.0	76.5	84.4	–	80.6
AdaBN	80.6	84.6	87.1	82.8	87.7	–	84.6	79.8	**87.9**	86.1	78.0	86.4	–	83.6
TENT	70.8	75.7	77.6	75.3	72.5	–	74.4	**83.6**	86.5	88.3	**78.5**	**87.4**	–	**84.9**
CoTTA	77.9	**85.3**	82.8	81.2	76.6	–	80.8	75.0	82.5	85.1	76.6	84.5	–	80.7
EATA	79.2	83.5	86.5	82.5	87.6	–	83.9	68.7	80.2	80.5	72.7	74.5	–	75.3
RoTTA	77.4	85.1	82.3	80.9	74.9	–	80.1	74.9	82.4	85.0	76.4	84.2	–	80.6
CoWA	**82.0**	84.6	**88.1**	**83.6**	**89.9**	–	**85.6**	83.4	83.6	**88.8**	78.2	87.1	–	84.2
	Target: NIH							**Target: CheXpert**						
No Adapt.	56.6	62.3	67.4	67.4	64.1	69.7	64.6	74.3	71.7	79.4	80.8	**71.2**	78.3	76.0
AdaBN	59.4	63.9	**71.5**	**67.7**	**69.4**	70.0	**67.0**	**80.2**	**79.9**	78.5	81.9	56.7	79.9	76.2
TENT	50.1	50.1	53.5	55.3	52.0	63.0	54.0	77.7	76.5	80.6	83.2	65.5	81.1	77.4
CoTTA	56.3	55.4	59.3	61.2	66.6	64.8	60.6	74.3	71.7	79.4	80.8	71.1	78.3	75.9
EATA	59.3	**67.0**	70.5	67.0	63.7	70.1	66.3	77.9	77.8	**80.9**	82.9	60.6	**81.3**	76.9
RoTTA	56.8	63.3	66.2	66.7	59.5	70.1	63.8	74.3	71.7	79.4	80.8	**71.2**	78.3	75.9
CoWA	**59.7**	66.8	71.1	67.3	65.4	**70.7**	66.8	79.2	78.1	80.5	**83.7**	66.1	79.6	**77.9**

others, consistently ranking first or second with only small gaps. In contrast, TENT and AdaBN show clear variability across domain pairs, performing well in some cases but degrading substantially in others. CoTTA and RoTTA often remain close to the unadapted baseline under larger shifts. Class-wise results further explain these trends. Competing methods frequently suffer marked drops on low-prevalence pathologies (e.g., Pneumothorax), where entropy minimization or simple BN updates can reinforce unreliable predictions. CoWA consistently alleviates such class-specific degradations, indicating that its higher mean AUROC is driven by preventing severe failures on vulnerable classes. In clinical settings, where a severe drop on any single pathology can compromise diagnosis, such worst-case robustness is as important as mean performance: notably, CoWA is the only method that avoids severe drops below the unadapted baseline across all shifts.

4.3 Analysis

We analyze CoWA from three perspectives: (1) the quality of the estimated co-occurrence matrix, (2) the reliability of the resulting sample weights, and (3) an ablation study on the norm choice in the weighting mechanism.

Does the Estimated Matrix Preserve Target Label Structure? CoWA assumes that the estimated co-occurrence matrix $\mathbf{M}$ captures the label dependency structure of the target domain during adaptation. To evaluate this assumption, we construct a reference matrix by passing target-domain test data through the same architecture trained with full supervision on the target domain, which serves as a proxy for the true target co-occurrence structure. Fig. 3 illustrates how $\mathbf{M}$ evolves under the CheXpert $\rightarrow$ MIMIC shift. At Batch 1, CoWA and the baseline are identical, as both models generate predictions from the source-trained model without adaptation. As adaptation proceeds, however, their behaviors diverge. With CoWA (top), the estimated matrix progressively aligns with the reference matrix. In contrast, the baseline, which does not update during testing, increasingly deviates from the reference in co-occurrence patterns involving Pneumothorax. In particular, its co-occurrence values between Pneumothorax and other pathologies become less distinguishable, resulting in blurred dependency structure. These results indicate that co-occurrence-guided weighting better preserves target-domain label dependencies during adaptation than the baseline.

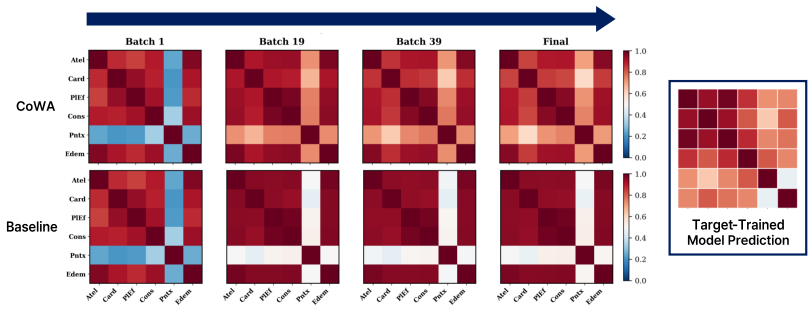

Fig. 3. Co-occurrence matrix evolution (CheXpert $\rightarrow$ MIMIC). CoWA (top) gradually aligns toward the target-trained reference (right), while the baseline (bottom) progressively loses clear dependency structure over batches.

Do CoWA Weights Reflect Sample Reliability? To examine whether CoWA weights reflect sample reliability, we analyze prediction quality as a function of weight under the CheXpert $\rightarrow$ MIMIC domain shift. Target samples are grouped into weight bins, and classification performance within each bin is computed using ground-truth labels. We report AUPRC due to the severe class imbalance

in chest X-ray datasets [14]. Fig. 4 presents the weight distribution from the final adaptation batch (blue) together with the corresponding performance for each bin (red). A positive association is observed between weight magnitude and predictive performance: lower-weight samples exhibit weaker predictive quality, whereas higher-weight samples achieve stronger performance. These findings suggest that co-occurrence consistency correlates with sample reliability, supporting the role of CoWA weights as an implicit reliability estimator during adaptation.

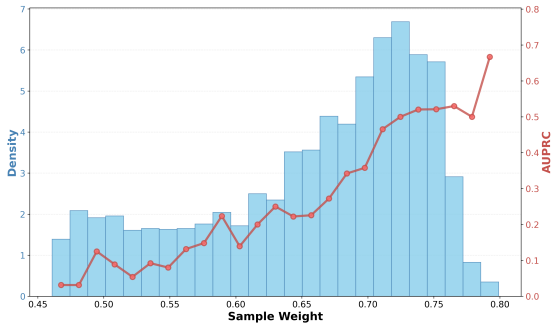

Fig. 4. Sample weight distribution and prediction reliability. Blue bars show the weight distribution after adaptation, and red line shows AUPRC per weight bin. Higher-weight samples consistently achieve better prediction quality.

Ablation Study. We first verify that the proposed reweighting stabilizes adaptation: the variance of the BN affine gradient ℓ_2-norm under CoWA is about half that of TENT (1.02×10^{-5} vs. 2.38×10^{-5}), confirming that consistency-based reweighting yields more stable updates. We then justify the choice of the Frobenius norm (ℓ_F) in Eq. 3 by comparing it against ℓ_1, ℓ_2, and ℓ_∞ alternatives over all six domain shift scenarios (Table 2). ℓ_F achieves the highest mean AUROC while also exhibiting the lowest variance across the six shifts, indicating more stable adaptation. Unlike ℓ_1, ℓ_2, or ℓ_∞, which can be dominated by large deviations in a few individual entries, ℓ_F aggregates squared differences over all entries and thus better captures the overall discrepancy of the co-occurrence matrix, making it more appropriate for aligning label dependencies rather than individual entries.

5 Conclusion

We propose CoWA, a TTA method that leverages disease co-occurrence patterns estimated from model predictions as a per-sample reliability signal for entropy minimization. Unlike standard TTA approaches that treat all samples equally without considering label dependencies, CoWA downweights predictions with inconsistent label combinations, enabling adaptation to focus on structurally con-

Table 2. AUROC of norm ablation in CoWA across six domain-shift scenarios.

Norm	Source: NIH			Source: CheXpert			Mean
	CheXpert	MIMIC	VinDr	NIH	MIMIC	VinDr	
ℓ_1	77.9	51.8	84.2	59.5	68.4	85.0	71.1 ($\pm$12.2)
ℓ_2	77.8	50.6	84.2	59.4	67.0	79.7	69.8 ($\pm$11.8)
ℓ_∞	77.9	50.6	84.2	59.4	68.4	83.2	70.6 ($\pm$12.3)
ℓ_F	**77.9**	**65.5**	**84.2**	**66.8**	**68.4**	**85.6**	**74.7 ($\pm$8.0)**

sistent samples. Experiments across multiple chest X-ray domain shift scenarios demonstrate consistent improvements over established baselines. By incorporating target-domain co-occurrence structure into TTA without requiring additional annotations or architectural changes, CoWA offers a simple and deployable solution for improving diagnostic robustness under domain shift. As the estimated patterns may be sensitive to dataset-level biases, extending CoWA to capture more robust, causally grounded relationships is a promising future direction.

Acknowledgments. This work was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) (Grant No. RS-2022-II220984) and in part by the National Research Foundation of Korea (NRF) (Grant No. RS-2024-00338859, RS-2025-00515481), both funded by the Ministry of Science and ICT (MSIT), Republic of Korea.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Cardinale, L., Priola, A.M., Moretti, F., Volpicelli, G.: Effectiveness of chest radiography, lung ultrasound and thoracic computed tomography in the diagnosis of congestive heart failure. World journal of radiology **6**(6), 230 (2014)
2. Cohen, J.P., Viviano, J.D., Bertin, P., Morrison, P., Torabian, P., Guarrera, M., Lungren, M.P., Chaudhari, A., Brooks, R., Hashir, M., Bertrand, H.: Torchxrayvision: A library of chest x-ray datasets and models. In: Proceedings of The 5th International Conference on Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 172, pp. 231–249. PMLR (2022)
3. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International Conference on Machine Learning. pp. 1180–1189. PMLR (2015)
4. Ghafoorian, M., Mehrtash, A., Kapur, T., Karssemeijer, N., Marchiori, E., Pesteie, M., Guttmann, C.R., De Leeuw, F.E., Tempany, C.M., Van Ginneken, B., 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. pp. 516–524. Springer (2017)
5. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: International Conference on Learning Representations (2019)
6. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph

dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019)
7. Johnson, A., Lungren, M., Peng, Y., Lu, Z., Mark, R., Berkowitz, S., Horng, S.: Mimic-cxr-jpg-chest radiographs with structured labels. PhysioNet **101**, 215–220 (2019)
8. Li, Y., Wang, N., Shi, J., Hou, X., Liu, J.: Adaptive batch normalization for practical domain adaptation. Pattern Recognition **80**, 109–117 (2018)
9. Milne, E., Pistolesi, M., Miniati, M., Giuntini, C.: The radiologic distinction of cardiogenic and noncardiogenic edema. American journal of roentgenology **144**(5), 879–894 (1985)
10. Musa, A., Ibrahim Adamu, M., Kakudi, H.A., Hernandez, M., Lawal, Y.: Analyzing cross-population domain shift in chest x-ray image classification and mitigating the gap with deep supervised domain adaptation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 585–595. Springer (2024)
11. Nguyen, H.Q., Lam, K., Le, L.T., Pham, H.H., Tran, D.Q., Nguyen, D.B., Le, D.D., Pham, C.M., Tong, H.T., Dinh, D.H., et al.: Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. Scientific Data **9**(1), 429 (2022)
12. Niu, S., Wu, J., Zhang, Y., Chen, Y., Zheng, S., Zhao, P., Tan, M.: Efficient test-time model adaptation without forgetting. In: Proceedings of the 39th International Conference on Machine Learning. pp. 16888–16905. PMLR (2022)
13. Pooch, E.H.P., Ballester, P.L., Barros, R.C.: Can we trust deep learning based diagnosis? the impact of domain shift in chest radiograph classification. In: Thoracic Image Analysis (TIA 2020), Lecture Notes in Computer Science. vol. 12502, pp. 74–83. Springer (2020)
14. Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. PloS one **10**(3), e0118432 (2015)
15. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: International Conference on Learning Representations (2021)
16. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7201–7211 (2022)
17. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2097–2106 (2017)
18. Yao, L., Poblenz, E., Dagunts, D., Covington, B., Bernard, D., Lyman, K.: Learning to diagnose from scratch by exploiting dependencies among labels. arXiv preprint arXiv:1710.10501 (2017)
19. Yuan, L., Xie, B., Li, S.: Robust test-time adaptation in dynamic scenarios. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15922–15932 (2023)
20. Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K.: Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Medicine **15**(11), e1002683 (2018)
21. Zhou, K., Yang, Y., Qiao, Y., Xiang, T.: Domain generalization with mixstyle. In: International Conference on Learning Representations (2021)