Towards Practical Algorithm Selection for Unsupervised Domain Adaptation in Medical Imaging

Yiheng Xiong[1], Luisa Gallée[1], Daniel Santak Wolf[1,2], Heiko Hillenhagen[1], and Michael Götz[1]

[1] Section of Experimental Radiology, Ulm University Medical Center, Germany
[2] Visual Computing Group, Ulm University, Germany
yiheng.xiong@uni-ulm.de

Abstract. Numerous unsupervised domain adaptation (UDA) algorithms exist, but for clinical practice, selecting the best-suited one along with proper hyperparameters often remains unclear, as the unlabeled deployment (target) domain prevents direct evaluation. We propose a label-free criterion that jointly selects the algorithm and hyperparameters for UDA. Given a pool of candidate models from multiple algorithms trained with different hyperparameters, our approach scores each candidate against an agreement reference, and selects the one with the highest score. The agreement reference is constructed in two levels without using target labels. First, we leverage multiple label-free selection signals, using each to nominate a model within every algorithm. Second, the nominated models are aggregated across algorithms to form a reference prediction for each unlabeled target sample. The candidate whose predictions agree most with this reference is then selected for deployment. Experimental results on four brain MRI and four chest X-ray datasets across seven clinically relevant transfer scenarios show that our method achieves better selection performance than other methods and remains effective across different algorithm pools. Our approach takes a step towards practical, label-free algorithm selection for clinical deployment of UDA. Our code is open sourced at Complete UDA Pipeline.

Keywords: Unsupervised Domain Adaptation · Algorithm Selection · Domain Shift · Deep Learning · Medical Imaging

1 Introduction

Unsupervised domain adaptation (UDA) aims to transfer knowledge from labeled source data to unlabeled target data, with recent algorithms showing robust adaptation performance on medical imaging benchmarks [18]. However, in clinical practice, selecting the best-suited algorithm often remains unclear, since the absence of target labels prevents direct evaluation. Existing unsupervised selection approaches largely consist of validators [25,27,33,29,23,30,12,32], label-free proxies that score each candidate checkpoint (a UDA model snapshot from

a training iteration) and select the best-scoring one, achieving reliable selection on general-vision UDA benchmarks such as VisDA [26]. However, two aspects remain less explored. First, each checkpoint is typically scored independently by a single validator, without exploiting the valuable agreement among different validators or among different candidate checkpoints. Hu et al. [13] combine predictions from validator-selected checkpoints within individual algorithms to leverage multiple validator signals, but do not exploit agreement across different algorithms. Second, while these methods are effective for hyperparameter selection given a fixed algorithm, the more realistic problem of joint algorithm and hyperparameter selection has received much less attention. Yang et al. [32] address algorithm selection by first choosing an algorithm under fixed hyperparameters and then selecting hyperparameters within it, but the joint selection of both, as encountered in clinical practice, remains largely unexplored.

To address this, given a pool of candidate checkpoints from multiple UDA algorithms trained with different hyperparameters, we propose a label-free selection criterion that scores each candidate against an agreement reference and selects the one with the highest score. The agreement reference is constructed in two levels. At the first level, we leverage multiple existing validators, using each to nominate a checkpoint within each algorithm; at the second level, the nominated checkpoints are aggregated across algorithms to form a reference prediction for each unlabeled target sample. The candidate whose predictions agree most with this reference is then selected for deployment.

We validate our method by selecting from multiple UDA algorithms across seven clinically relevant transfer scenarios on four brain MRI and four chest X-ray (CXR) datasets. Experimental results show that our method achieves better selection results than individual validators across all scenarios. Compared to the best individual validator, it also halves the target performance gap to the best available checkpoint. Our approach also remains effective under different algorithm pools and checkpoint densities. We summarize our contributions as:

- We propose a label-free criterion for joint algorithm and hyperparameter selection in UDA for medical imaging, which to our knowledge is the first to address this joint problem. Each candidate checkpoint is scored against an agreement reference to identify a single high-performing one across different UDA algorithms, supporting more reliable algorithm selection for clinical deployment where target labels are unavailable.
- We design a two-level construction of this agreement reference, in which validators select checkpoints within each algorithm at the first level, and their agreement across algorithms forms the reference at the second level, yielding a robust selection signal across diverse algorithm pools.

2 Methodology

2.1 Preliminary

UDA & Algorithm Selection in UDA. In UDA, a labeled source domain $\mathcal{D}_s = \{(x_i^s, y_i^s)\}$ and an unlabeled target domain $\mathcal{D}_t = \{x_j^t\}$ are given, and the

goal is to train a model using both $\mathcal{D}_s$ and $\mathcal{D}_t$ to make predictions on $\mathcal{D}_t$. We focus on UDA for classification, where a UDA algorithm is trained by jointly minimizing a supervised classification loss $\mathcal{L}_{\text{cls}}$ on the source domain and an adaptation loss $\mathcal{L}_{\text{adapt}}$ that aligns the two domains,

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{adapt}}, \tag{1}$$

where λ controls the adaptation strength.

In practice, many UDA algorithms exist, each with different $\mathcal{L}_{\text{adapt}}$ and typically trained under several hyperparameter configurations (such as different values of λ), with model snapshots saved at multiple iterations during training. We refer to each such snapshot as a checkpoint, so that a checkpoint is uniquely identified by an algorithm, a hyperparameter configuration, and a training iteration. Let $\mathcal{A} = \{a_1, \ldots, a_M\}$ denote a set of UDA algorithms, and let Θ denote the full pool of resulting checkpoints across all algorithms, hyperparameter configurations, and training iterations. Since target labels are unavailable, the target performance of any checkpoint $\theta \in \Theta$ cannot be measured directly, and it is therefore unknown which checkpoint performs best on $\mathcal{D}_t$. The task of algorithm selection is to identify a single checkpoint $\hat{\theta} \in \Theta$ for deployment, where the candidates span different UDA algorithms, hyperparameter configurations, and training iterations rather than being restricted to a single algorithm.

Validators. A validator is a function v that maps a checkpoint θ to a scalar score $v(\theta) \in \mathbb{R}$, computed without target labels, and serves as a proxy for target performance, so that the checkpoint with the best score is selected. The set of validators $\mathcal{V} = \{v_1, \ldots, v_N\}$ can roughly be divided into two categories. Source-guided includes Source-Risk [8], IWCV [29], DEV [33], and DEV-N [25]. Target-specific includes Entropy [23], InfoMax [24], MCC (V) [16], BNM (V) [25], Corr-C [30], SND [27], Class-AMI [25], MixVal [12] and TransScore [32].

2.2 Two-Level Agreement Reference Construction

Since target labels are unavailable, no direct reference exists for comparing checkpoints. We therefore construct a reference from the candidates themselves, in two levels, as illustrated in Figure 1. Each checkpoint $\theta \in \Theta$ produces a prediction $\hat{y}_j^t(\theta) = \arg\max_c p_c(x_j^t; \theta)$ for every target sample x_j^t, where $p_c(\cdot; \theta)$ denotes the predicted probability of class c.

Level 1: within-algorithm selection. For each algorithm $a \in \mathcal{A}$, every validator $v \in \mathcal{V}$ scores the checkpoints of a and nominates the best one,

$$\theta_{a,v} = \arg\max_{\theta \in \Theta_a} v(\theta), \tag{2}$$

where $\Theta_a = \{\theta_{a,k} \mid k = 1, \ldots, K\}$ denotes the checkpoints of algorithm a. This yields a set of validator-nominated checkpoints $\{\theta_{a,v} \mid v \in \mathcal{V}\}$ for each algorithm, where each checkpoint reflects the unique selection signal of a validator.

Level 2: across-algorithm agreement. The validator-nominated checkpoints across algorithms, $\{\theta_{a,v} \mid a \in \mathcal{A}, v \in \mathcal{V}\}$, are then combined by majority vote

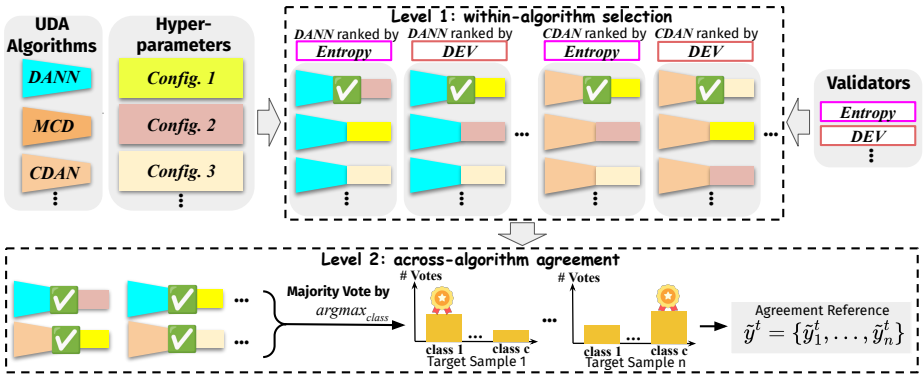

Fig. 1: Two-level agreement reference construction. Each candidate combines an algorithm (trapezoid) and a hyperparameter configuration (rectangle). At Level 1, each validator nominates the top checkpoint within each algorithm (green check). At Level 2, the nominated checkpoints vote per target sample, and the majority class (medal) forms the agreement reference $\tilde{y}^t$.

across their predictions. For each target sample x_j^t, the agreement reference is

$$\tilde{y}_j^t = \arg\max_c \sum_{a \in \mathcal{A}} \sum_{v \in \mathcal{V}} \mathbb{1}\left[\hat{y}_j^t(\theta_{a,v}) = c\right], \tag{3}$$

where $\mathbb{1}[\cdot]$ is the indicator function. By selecting within each algorithm using each validator and then aggregating across algorithms, the reference draws on the complementary strengths of different UDA algorithms and validators, rather than relying on any single one.

2.3 Label-Free Selection Criterion

Given the agreement reference $\tilde{y}^t = \{\tilde{y}_j^t\}_{j=1}^{|\mathcal{D}_t|}$ as a label-free target, every candidate checkpoint in the pool Θ is scored by how closely its predictions match the reference $\tilde{y}^t$. The agreement score of a checkpoint θ is measured as the per-class fraction of target samples on which its prediction coincides with the reference, averaged over all classes,

$$s(\theta) = \frac{1}{C} \sum_{c=1}^{C} \frac{\sum_j \mathbb{1}\left[\hat{y}_j^t(\theta) = c\right] \mathbb{1}\left[\tilde{y}_j^t = c\right]}{\sum_j \mathbb{1}\left[\tilde{y}_j^t = c\right]}, \tag{4}$$

where C is the number of classes. The single checkpoint with the highest agreement score is selected for deployment,

$$\hat{\theta} = \arg\max_{\theta \in \Theta} s(\theta). \tag{5}$$

The selected $\hat{\theta}$ is a single checkpoint from one UDA algorithm, obtained without target labels. The agreement reference is used only to identify it, without the need to deploy all the nominated checkpoints.

3 Experiments

Datasets. To construct medical UDA scenarios, four widely adopted brain MRI datasets are used: ADNI-1, ADNI-2, ADNI-3 [15], and AIBL [6]. In addition, four publicly available CXR datasets are used: RSNA [31], Child CXR [17], LDD [2], and CRD [1]. Each dataset is treated as a separate domain, and transfer is performed within each modality. The brain MRI datasets contain Alzheimer's disease and cognitively normal subjects, while the CXR datasets contain pneumonia and non-pneumonia subjects. Both modalities undergo standard preprocessing following prior work [10,34]. Dataset statistics are summarized in Table 1.

Table 1: Dataset statistics for brain MRI and CXR datasets. Values are the number of samples per class. AD: Alzheimer's disease; CN: cognitively normal; Pneu.: pneumonia; Non-Pneu.: non-pneumonia.

Brain MRI		**CXR**	
Dataset	**AD / CN**	**Dataset**	**Pneu. / Non-Pneu.**
ADNI-1 [15]	200 / 221	RSNA [31]	6,012 / 20,672
ADNI-2 [15]	159 / 232	Child CXR [17]	4,273 / 1,583
ADNI-3 [15]	85 / 431	LDD [2]	5,776 / 3,919
AIBL [6]	78 / 477	CRD [1]	9,237 / 10,319

UDA Algorithms. Algorithm selection is performed over a pool of ten established UDA algorithms spanning multiple paradigms: feature-distance minimization (MMD [21]), adversarial alignment (DANN [9], CDAN [22], and DALN [3]), information maximization (MCC [16]), SVD loss (BNM [4]), pseudo labeling (ATDOC [19]), classifier discrepancy (MCD [28]), and medical-specific AD2A [10] (brain MRI) and CoUDA [34] (CXR).

Experimental Setup. Following Guan et al. [10], five UDA scenarios (source→target) are constructed for brain MRI: ADNI-1→ADNI-2, ADNI-1→ADNI-3, ADNI-2→ADNI-1, ADNI-2→ADNI-3, and ADNI-1+2→AIBL. Following Feng et al. [7] and Liu et al. [20], two scenarios are constructed for CXR: RSNA→Child CXR and LDD→CRD. Following Musgrave et al. [24], both source and target data are split into training and validation sets, and target performance is measured on the target validation set using balanced accuracy. Stratified five-fold cross-validation is performed, and the mean ± standard deviation is reported. We do not hold out a separate target test set, as the selection procedure is label-free: given any new unlabeled target data, the validation scores can be recomputed directly on it, and the best-scored checkpoint selected.

Implementation Details. For all experiments, the classification head is a three-layer MLP with a dropout rate of 0.5. For brain MRI, a 3D ResNet-50 [11] trained from scratch is used as the backbone, with a batch size of eight per domain. For CXR, a 2D ResNet-50 pretrained on ImageNet [5] is used for RSNA→Child CXR and a pretrained 2D DenseNet-121 [14] for LDD→CRD, both with a batch size of 48 per domain. The two backbones let us assess

whether our selection criterion remains robust under a different architecture. All algorithms are trained with AdamW (weight decay 1e-4) and a one-cycle learning rate schedule with warm-up and a peak learning rate of 1e-3, for 10k iterations on brain MRI and 30k iterations on CXR. The adaptation strength is varied over $\lambda \in \{0.1, 0.5, 1.0\}$, giving three runs per algorithm. After warm-up, checkpoints are saved at uniform intervals, yielding 50 checkpoints per run and 150 checkpoints per algorithm. $\mathcal{L}_{\text{cls}}$ is cross-entropy with class-balanced weighting derived from the source labels. All runs are conducted on an A6000 GPU. Training an algorithm for one fold and one value of λ takes roughly 6 to 15 hours on brain MRI and 2 to 5 hours on CXR, depending on the algorithm. The output logits of each checkpoint are saved during training, so running the validators requires no additional forward passes.

4 Results and Discussion

Main Results. Table 2 reports the selection results on all UDA scenarios. Due to space constraints, we show the top five individual validators ranked by overall average performance. Our method achieves the highest overall average accuracy (86.3%), improving over the best individual validator (Class-AMI, 81.0%) by 5.3%, and reduces the gap to the *Oracle*, the best checkpoint identifiable with target labels, from 10.4% to 5.1%. The improvement is consistent across individual scenarios: our method ranks first on all five brain MRI scenarios and both CXR scenarios. On CXR, where the two scenarios use different backbones, our method also obtains the best average, indicating that the selection criterion remains effective under a different architecture.

Table 2: Selection results across UDA scenarios on brain MRI and CXR, reported as target accuracy (%). *Oracle* selects the best checkpoint using target labels and serves as an upper bound. **Bold** indicates the best non-*Oracle* result. Our method achieves the highest average accuracy and the smallest gap to *Oracle*.

	Brain MRI Scenarios						**CXR Scenarios**			**Overall**	
Validator	A1→A2	A1→A3	A2→A1	A2→A3	A1+2→AIBL	**Avg.**	RSNA→Child	LDD→CRD	**Avg.**	All	*Δ Oracle*↓
Oracle	93.8 ±2.1	94.2 ±3.1	92.5 ±1.8	92.1 ±2.4	93.8 ±2.1	93.3	89.5 ±0.56	83.8 ±0.69	86.7	91.4	0.0
InfoMax	83.2 ±7.2	70.6 ±7.9	85.7 ±2.9	80.6 ±3.2	81.6 ±5.0	80.3	78.8 ±5.9	80.5 ±1.7	79.6	80.1	11.3
Source-Risk	83.5 ±2.9	82.9 ±5.2	79.5 ±6.3	82.3 ±7.7	84.0 ±5.1	82.4	73.2 ±4.3	77.3 ±1.6	75.2	80.4	11.0
DEV-N	85.1 ±1.3	84.5 ±5.0	79.5 ±6.3	78.3 ±9.7	82.7 ±5.0	82.0	75.8 ±3.3	77.4 ±1.4	76.6	80.5	10.9
DEV	79.4 ±13	84.2 ±7.8	84.9 ±3.3	80.3 ±7.8	89.0 ±1.7	83.6	72.6 ±12	75.5 ±7.9	74.1	80.9	10.5
Class-AMI	80.8 ±3.6	79.2 ±9.7	81.8 ±3.0	85.0 ±3.7	88.0 ±3.0	83.0	74.2 ±8.7	78.0 ±1.8	76.1	81.0	10.4
Ours	**89.0** ±4.6	**88.1** ±3.3	**88.9** ±4.0	**86.1** ±6.1	**90.2** ±3.8	**88.5**	**80.5** ±2.0	**81.3** ±0.85	**80.9**	**86.3**	**5.1**

Ablation Studies. We ablate three aspects of our method: the aggregation structure underlying the reference construction; the algorithm pool, varying both its size and composition; and the number of checkpoints saved per algorithm.
Aggregation structure. Table 3 compares our two-level design with two flat baselines. The first, *all checkpoints*, builds the reference by majority vote over every

Table 3: Ablation on aggregation structure, reported as selected target accuracy (%). *All checkpoints* builds the reference from every checkpoint with no validator nomination or algorithm structure; *Per-validator* retains validator nomination but ignores algorithm boundaries. Our two-level design achieves the highest overall accuracy and the smallest gap to the *Oracle*.

	Brain MRI Scenarios						CXR Scenarios			Overall	
Aggregation	A1→A2	A1→A3	A2→A1	A2→A3	A1+2→AIBL	**Avg.**	RSNA→Child	LDD→CRD	**Avg.**	All	*ΔOracle*↓
All checkpoints	83.3±6.3	83.1±4.3	81.8±4.4	82.0±6.0	86.8±4.6	83.4	80.4±2.6	79.5±0.7	80.0	82.4	9.0
Per-validator	86.0±2.5	80.5±6.1	79.8±7.7	81.9±4.9	83.5±6.5	82.3	79.1±3.0	79.8±2.8	79.4	81.5	9.9
Two-level (Ours)	**89.0**±4.6	**88.1**±3.3	**88.9**±4.0	**86.1**±6.1	**90.2**±3.8	**88.5**	**80.5**±2.0	**81.3**±0.9	**80.9**	**86.3**	**5.1**

checkpoint, with no validator nomination or algorithm structure. The second, *per-validator*, retains validator nomination but ignores algorithm boundaries by pooling all checkpoints into a single global pool. Our two-level design achieves better selection performance than both flat variants, suggesting its effectiveness in producing more reliable agreement reference.

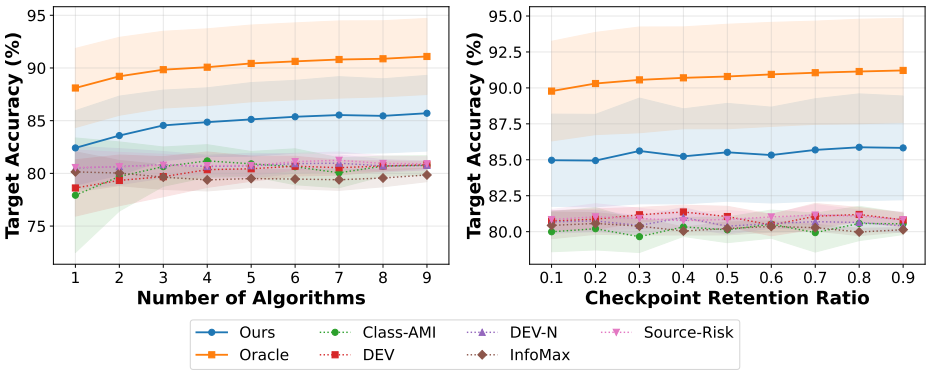

Fig. 2: Ablation on the algorithm pool (left) and the fraction of checkpoints retained per algorithm (right), reported as target accuracy (%) averaged over all UDA scenarios, with shaded regions indicating standard deviation. At each pool size, subsets are sampled randomly, varying both the number and composition of algorithms. Our method maintains a stable gap to *Oracle* and stays above other individual validators across both axes.

Algorithm pool. Our method builds the agreement reference from checkpoints across multiple algorithms, so a natural question is whether it remains robust under different algorithm pools. To test this, we vary the algorithm pool by randomly sampling several subsets at each size from one to all nine available per scenario (eight general plus one modality-specific), reporting the average over the sampled subsets and all UDA scenarios. Since the subsets are sampled randomly, this varies both the number and the composition of algorithms in the pool. As

shown in Figure 2, compared to individual validators, our method maintains a smaller and more stable gap to the *Oracle* across different pools, suggesting that its selection quality holds up well as the pool varies.

Number of checkpoints per algorithm. Our method draws on the checkpoints saved during training, so we also test whether it depends on having many checkpoints per algorithm. We subsample each algorithm's checkpoints at ratios from 0.1 to 0.9, repeating the sampling ten times at each ratio and reporting the average over the sampled draws and all UDA scenarios. As shown in Figure 2, our method keeps a stable gap to the *Oracle* and remains the highest among other validators across different subsample ratios. This suggests that our approach is able to remain effective with relatively few checkpoints per algorithm.

Discussion. To apply UDA in clinical practice, selecting the best-suited algorithm is particularly difficult: target performance cannot be evaluated directly, and the candidate pool is heterogeneous, spanning not just different hyperparameters but different algorithms. Our two-level agreement reference is designed to address this by drawing on multiple validators within each algorithm and then aggregating across diverse algorithms, with the aim of combining complementary signals rather than relying on any single one. In Table 3, we observe that *All checkpoints* outperforms *Per-validator* on average. A possible reason is that, without algorithm boundaries, a validator may nominate checkpoints mainly from algorithms whose training objective is similar to its own criterion. This could reduce the diversity of the reference, whereas using all checkpoints preserves it. Although the reference is built from multiple checkpoints, it is used only to identify a single checkpoint for deployment; the nominated checkpoints do not need to be stored or deployed. Beyond this, the selection quality of our method remains stable across different algorithm pools. We interpret this through two regimes. With few algorithms, the selection task is easier since there are fewer candidates, and the first level of agreement is often enough to identify a good checkpoint. With many algorithms, the task is harder as the candidate space is more heterogeneous, but the larger pool gives the second level of agreement more diversity to draw on, making the reference more reliable. From a clinical deployment perspective, since the best-available checkpoint improves as more algorithms are added, we recommend training multiple UDA algorithms spanning different paradigms when resources allow, giving the method stronger options to select among; when resources are limited, the method still selects reliably from a smaller pool, though deployed performance may be bounded by it.

Limitations and Future Work. While our method achieves better selection than individual validators, a gap to the *Oracle* remains, most notably on RSNA→Child CXR. Selecting a single strong checkpoint also requires training potentially multiple algorithms, which is computationally costly; improving the efficiency of this process is an important direction for future work. In addition, our study is limited to binary classification and uses balanced accuracy as the selection target; extending to multi-class classification or segmentation, and to other clinically relevant metrics such as sensitivity, are natural next steps.

5 Conclusion

In this paper, we propose, to our knowledge, the first label-free criterion for joint algorithm and hyperparameter selection in UDA for medical imaging. Our criterion scores each candidate against a two-level agreement reference, built from multiple validators within each algorithm and aggregated across algorithms. The highest-scoring candidate is then selected for deployment. Experimental results on four brain MRI and four CXR datasets across seven clinically relevant transfer scenarios show that our method achieves better selection results than individual validators. Our approach also remains effective across different algorithm pools and checkpoint densities. We see this as a step towards practical, label-free algorithm selection for clinical deployment of UDA.

Prospects of Application. A common scenario could be adapting from labeled data at hospital A to unlabeled data at hospital B, which may differ in scanner, acquisition protocol, or patient population. By applying our method, a practitioner could jointly select the algorithm and hyperparameters from many candidates, obtaining a more reliable deployable model to support physicians in analyzing the data at hospital B.

Acknowledgments. This study was funded by the German Research Foundation DFG (Project: KEMAI, GRK 3012 – 520750254) and by the German Federal Ministry of Research, Technology and Space BMFTR as part of the University Medicine Network 3.0 (Project: RACOON, 01KX2524).

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

References

1. Covid-19 radiography database. https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database
2. Lungs disease dataset (4 types). https://www.kaggle.com/datasets/omkarmanohardalvi/lungs-disease-dataset-4-types
3. Chen, L., Chen, H., Wei, Z., et al.: Reusing the task-specific classifier as a discriminator: Discriminator-free adversarial domain adaptation. In: CVPR. pp. 7181–7190 (2022)
4. Cui, S., Wang, S., Zhuo, J., et al.: Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In: CVPR. pp. 3941–3950 (2020)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
6. Ellis, K.A., Bush, A.I., Darby, D., et al.: The australian imaging, biomarkers and lifestyle (aibl) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of alzheimer's disease. International psychogeriatrics **21**(4), 672–687 (2009)

7. Feng, Y., Wang, Z., Xu, X., Wang, Y., Fu, H., Li, S., Zhen, L., Lei, X., Cui, Y., Ting, J.S.Z., et al.: Contrastive domain adaptation with consistency match for automated pneumonia diagnosis. Medical Image Analysis **83**, 102664 (2023)
8. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: ICML. pp. 1180–1189. PMLR (2015)
9. Ganin, Y., Ustinova, E., Ajakan, H., et al.: Domain-adversarial training of neural networks. Journal of machine learning research **17**(59), 1–35 (2016)
10. Guan, H., Liu, Y., Yang, E., et al.: Multi-site mri harmonization via attention-guided deep domain adaptation for brain disorder identification. Medical image analysis **71**, 102076 (2021)
11. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
12. Hu, D., Liang, J., Liew, J.H., Xue, C., Bai, S., Wang, X.: Mixed samples as probes for unsupervised model selection in domain adaptation. Advances in Neural Information Processing Systems **36**, 37923–37941 (2023)
13. Hu, D., Luo, R., Liang, J., et al.: Towards reliable model selection for unsupervised domain adaptation: An empirical study and a certified baseline. NeurIPS **37**, 135883–135903 (2024)
14. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
15. Jack Jr, C.R., Bernstein, M.A., Fox, N.C., et al.: The alzheimer's disease neuroimaging initiative (adni): Mri methods. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine **27**(4), 685–691 (2008)
16. Jin, Y., Wang, X., Long, M., et al.: Minimum class confusion for versatile domain adaptation. In: ECCV. pp. 464–480. Springer (2020)
17. Kermany, D.S., Goldbaum, M., Cai, W., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. cell **172**(5), 1122–1131 (2018)
18. Kumari, S., Singh, P.: Deep learning for unsupervised domain adaptation in medical imaging: Recent advancements and future perspectives. Computers in Biology and Medicine **170**, 107912 (2024)
19. Liang, J., Hu, D., Feng, J.: Domain adaptation with auxiliary target domain-oriented classifier. In: CVPR. pp. 16632–16642 (2021)
20. Liu, W., Ni, Z., Chen, Q., Ni, L.: Attention-guided partial domain adaptation for automated pneumonia diagnosis from chest x-ray images. IEEE Journal of Biomedical and Health Informatics **27**(12), 5848–5859 (2023)
21. Long, M., Cao, Y., Wang, J., et al.: Learning transferable features with deep adaptation networks. In: ICML. pp. 97–105. PMLR (2015)
22. Long, M., Cao, Z., Wang, J., et al.: Conditional adversarial domain adaptation. NeurIPS **31** (2018)
23. Morerio, P., Cavazza, J., Murino, V.: Minimal-entropy correlation alignment for unsupervised deep domain adaptation. arXiv preprint arXiv:1711.10288 (2017)
24. Musgrave, K., Belongie, S., Lim, S.N.: Unsupervised domain adaptation: A reality check. arXiv preprint arXiv:2111.15672 (2021)
25. Musgrave, K., Belongie, S., Lim, S.N.: Three new validators and a large-scale benchmark ranking for unsupervised domain adaptation. arXiv preprint arXiv:2208.07360 (2022)
26. Peng, X., Usman, B., Kaushik, N., Hoffman, J., Wang, D., Saenko, K.: Visda: The visual domain adaptation challenge. arXiv preprint arXiv:1710.06924 (2017)

27. Saito, K., Kim, D., Teterwak, P., et al.: Tune it the right way: Unsupervised validation of domain adaptation via soft neighborhood density. In: ICCV. pp. 9184–9193 (2021)
28. Saito, K., Watanabe, K., Ushiku, Y., et al.: Maximum classifier discrepancy for unsupervised domain adaptation. In: CVPR. pp. 3723–3732 (2018)
29. Sugiyama, M., Krauledat, M., Müller, K.R.: Covariate shift adaptation by importance weighted cross validation. JMLR **8**(5) (2007)
30. Tu, W., Deng, W., Gedeon, T., et al.: Assessing model out-of-distribution generalization with softmax prediction probability baselines and a correlation method
31. Wang, X., Peng, Y., Lu, L., et al.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: CVPR. pp. 2097–2106 (2017)
32. Yang, J., Qian, H., Xu, Y., Wang, K., Xie, L.: Can we evaluate domain adaptation models without target-domain labels? In: International Conference on Learning Representations. vol. 2024, pp. 35061–35081 (2024)
33. You, K., Wang, X., Long, M., et al.: Towards accurate model selection in deep unsupervised domain adaptation. In: ICML. pp. 7124–7133. PMLR (2019)
34. Zhang, Y., Wei, Y., Wu, Q., et al.: Collaborative unsupervised domain adaptation for medical image diagnosis. IEEE TIP **29**, 7834–7844 (2020)