Investigating Test-Time Adaptation of Convolutional Neural Networks for Medical Image Analysis under Distribution Shifts

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Deep learning models often fail to generalize when classifying medical images that originate from a different data distribution than the training images. Their error rate increases, posing a severe risk to patients. Test-Time Training (TTT) with Masked Autoencoders (MAE) addresses distribution shifts by adapting a trained model to unlabeled images at inference time, using a self-supervised objective. However, we find that its success in the natural image domain does not immediately transfer to the medical domain. A key limitation is that TTT can degrade model performance when the auxiliary loss used for inference-time adaptation is uncorrelated with the main task. Thus, we tackle distribution shifts from a different angle, eliminating the need for a suitable loss function or weight adaptation during inference time. We propose WhiCo, a novel Test-Time Adaptation method that mitigates distribution shifts by aligning test features with the training feature distribution through whitening and coloring. We evaluate WhiCo on diverse medical tasks, including COVID-19 diagnosis, brain hemorrhage detection, and organ identification. WhiCo improves accuracy by 3.5 to 8.4 percentage points on COVID-19 diagnosis, 3.4 to 10.3 points on brain hemorrhage detection, and by up to 21.9 points on organ identification. Moreover, WhiCo is model-agnostic and requires no expert knowledge to choose test-time hyperparameters. All code and data will be released upon acceptance.

Keywords: Test-Time Adaptation · Distribution Shift · CT.

1 Introduction

Deep learning models achieve expert-level performance in detecting Alzheimer's disease [13,14,19], COVID-19 [2,3,23], or cancer [6,12,38]. Therefore, they support radiologists in making reliable diagnoses while reducing errors caused by fatigue and capacity overload. Despite remarkable progress, models often fail to generalize under distribution shifts, i.e., when the source (training) and target (inference) data come from different data distributions [7,20]. In clinical practice, models inevitably encounter data distributions that differ from the source distribution due to varying image acquisition [18], disease prevalence [35], or patient demographics [1]. Decreasing model performance under distribution shifts

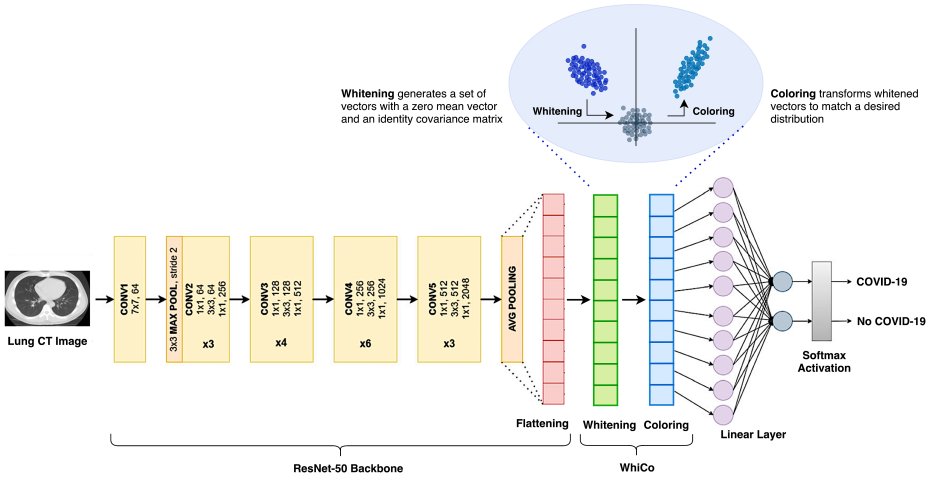

Fig. 1. A ResNet-50 backbone extracts feature vectors from CT scans. Whitening and coloring align the features with the source data distribution on which the model was trained to mitigate a distribution shift. The final layer predicts a label based on the aligned feature vectors. Figures adapted from Roy et al. [21] and Shazia et al. [22].

hinders real-world deployment, as misdiagnoses have severe consequences.

Unsupervised Domain Adaptation (UDA) and Test-Time Training (TTT) propose varying strategies to improve robustness to distribution shifts. UDA jointly trains a model on labeled source data and unlabeled target data to learn domain-invariant features, enabling generalization to a target domain without inference-time adaptation. To align feature distributions, UDA methods leverage self-supervised learning (SSL) [25,34], adversarial learning [5,28], or whitening and coloring transformation [21,24]. Despite being promising solutions, UDA is incompatible with real-world medical workflows because it trains a model on both source and target data before deployment. It requires that hospitals share training data from the target domain with model developers, which is often infeasible due to privacy regulations. In contrast to UDA, TTT first trains a model with labeled source data. At inference time, it adapts the model to unlabeled target data by leveraging an SSL loss [9,15,26] or entropy minimization [29,32,37]. For example, TTT-MAE [9] adapts a Vision Transformer (ViT) [8] via masked autoencoding and thereby greatly improves the model's robustness to distribution shifts in the natural image domain. Motivated by its success, we investigate TTT-MAE's applicability in the medical image domain. For an effective transfer, two modifications are necessary. At first, we replace the ViT with a ResNet-50 [11]. Consequently, we substitute the ViT-specific MAE with SparK [27], an MAE designed for ResNets. Unlike ResNets, ViTs lack spatial inductive biases and require massive annotated datasets to compensate for this [8]. In medical imaging, such datasets are challenging to collect due to the need for expert labels, ethical approval, and the low prevalence of rare diseases.

Pre-training on large datasets does not resolve this, as the bottleneck persists during downstream fine-tuning [10]. Moreover, convolutional networks can outperform ViTs [16]. We evaluate TTT-SparK on CT image classification, including organ identification, COVID-19 diagnosis, and brain hemorrhage detection, whereby we induce distribution shifts through CT scan corruptions [7,36]. Our results show that TTT-SparK does not improve the model's robustness to distribution shifts in these medical tasks. A limitation of TTT is that it can fail to prevent performance degradation under distribution shifts if the loss function used to adapt the model to the target distribution is uninformative for the task that the model solves after TTT [15]. These insights lead us to a different strategy to mitigate distribution shifts. We develop **WhiCo**, a novel Test-Time Adaptation method that leverages **Whi**tening and **Co**loring to align the target features with the source feature distribution (see Fig. 1). In contrast to UDA, it adapts a model to target data in-house at a clinical institution. WhiCo operates solely on target features without accessing source features at inference time, making it privacy-preserving. In contrast to TTT, WhiCo eliminates the need for a suitable loss to update model weights. We evaluate WhiCo on the same tasks and corruptions as TTT-SparK. In contrast to TTT-SparK, WhiCo increases the accuracy for most corruptions across all tasks. Specifically, it improves accuracy by up to 21.9 percentage points on organ identification, by 3.5 to 8.4 points on COVID-19 diagnosis, and by 3.4 to 10.3 points on brain hemorrhage detection. These results indicate that feature alignment is a valuable alternative to weight optimization for robust medical image analysis under distribution shifts.

2 Methods

2.1 Test-Time Training with SparK

Pre-Training. In line with TTT-MAE [9], we utilize a ResNet-50[1] that is pre-trained with SparK [27] on lung CT scans [4].
Training-Time Training. We fine-tune the pre-trained ResNet-50 on a downstream task with the approach described by Wolf et al. [30].
Test-Time Training. We employ the SSL approach *SparK* [27] as a TTT method. TTT-SparK leverages masked autoencoding to adapt a model to unlabeled target images. At first, TTT-SparK randomly masks patches of a target image. The remaining unmasked patches form a sparse image that an MAE takes as input. The MAE consists of an encoder that extracts features from a sparse image and a decoder that reconstructs the original image from these features (for details of the encoding-decoding procedure, see SparK [27]). Our fine-tuned ResNet serves as the encoder. A UNet-like module developed by Tian et al. [27] serves as the decoder. TTT-SparK backpropagates a self-supervised L^2-loss to train this MAE. This loss determines the error between the reconstructed pixels and the original unmasked pixels, which represent the labels.

[1] model downloaded from GitHub

Formally, the optimization objective takes the form of

$$f, g = argmin_{f,g} L_s(g \circ f(mask(\mathbf{x})), \mathbf{x}) \tag{1}$$

with the encoder f, the decoder g, the target image $\mathbf{x}$, and the reconstruction L^2-loss L_s. Through this optimization, TTT adapts the ResNet to target images that may come from a different data distribution than the training images.
Classification. After TTT, the adjusted ResNet predicts a label for the unmasked target image $\mathbf{x}$.

2.2 Test-Time Adaptation with WhiCo

Pre-Training. We take the same pre-trained ResNet-50 as for TTT-SparK to facilitate comparability in the experiments.
Training-Time Training. We fine-tune the pre-trained ResNet on a downstream task with the procedure of Wolf et al. [30]. After fine-tuning, we pass each training image through the ResNet backbone (i.e., all layers except the last layer) and collect the extracted feature vector. We calculate summary statistics, i.e., the mean vector μ_{train} and the covariance matrix $\mathbf{\Sigma}_{train}$, of the collected set of feature vectors. To decouple the training and inference phases, we store these summary statistics. WhiCo only accesses the fine-tuned model and summary statistics without inquiring the training data at inference time. The reason for decoupling is that when deploying the model in a clinical institution, data privacy regulations often prohibit sharing the training data. However, the summary statistics neither directly reveal nor permit inference of sensitive information about patients in the training cohort. Thus, summary statistics are a valuable means to give information about the training feature distribution without violating privacy regulations.
Test-Time Adaptation. Given a batch of test images $B_{images} = \{I_1, I_2, ..., I_n\}$, we pass each test image $I_j \in B_{images}$ through the ResNet backbone and collect the resulting feature vector $\mathbf{x}_j$. Then, we utilize the gathered batch of feature vectors $B_{features} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ to calculate summary statistics, i.e., the mean vector μ_{test} and the covariance matrix $\mathbf{\Sigma}_{test}$. Given the covariance matrices $\mathbf{\Sigma}_{train}$ and $\mathbf{\Sigma}_{test}$, we leverage the eigendecomposition to compute the (inverse) square root covariance matrices

$$\mathbf{\Sigma}_{train}^{\frac{1}{2}} = \mathbf{E}_{train}\ \mathbf{\Lambda}_{train}^{\frac{1}{2}}\ \mathbf{E}_{train}^{T} \tag{2}$$

$$\mathbf{\Sigma}_{test}^{-\frac{1}{2}} = \mathbf{E}_{test}\ \mathbf{\Lambda}_{test}^{-\frac{1}{2}}\ \mathbf{E}_{test}^{T}. \tag{3}$$

The variables $\mathbf{E}_{train}$ and $\mathbf{E}_{test}$ denote the orthogonal matrices whose columns contain the orthonormal eigenvectors of $\mathbf{\Sigma}_{train}$ and $\mathbf{\Sigma}_{test}$, respectively. The diagonal matrix $\mathbf{\Lambda}_{train}^{\frac{1}{2}}$ contains the square root of the eigenvalues from $\mathbf{\Sigma}_{train}$ as entries. Similarly, the diagonal matrix $\mathbf{\Lambda}_{test}^{-\frac{1}{2}}$ stores the inverse square root of the eigenvalues from $\mathbf{\Sigma}_{test}$. These matrices store the eigenvalues in descending order and arrange the eigenvectors according to their corresponding eigenvalues.

We add a small constant ϵ = 1e-5 if an eigenvalue is zero to ensure the matrix is invertible (see the matrix cookbook [17] for mathematical background).
These matrices are necessary for the whitening and coloring steps. Fig. 1 illustrates the model inference sequence extended by whitening and coloring. At first, we whiten each feature vector $\mathbf{x}_j \in B_{features}$, extracted by the backbone, with

$$\mathbf{x}_j^{whi} = \mathbf{\Sigma}_{test}^{-\frac{1}{2}}(\mathbf{x}_j - \mu_{test}). \tag{4}$$

Whitening a set of vectors yields a transformed vector set whose mean is the zero vector and whose covariance matrix is the identity matrix. Coloring transforms a whitened vector to match a desired distribution (in our case, the source distribution). Therefore, we apply coloring to each whitened feature vector $\mathbf{x}_j^{whi}$

$$\mathbf{x}_j^{whico} = \mathbf{\Sigma}_{train}^{\frac{1}{2}}\ \mathbf{x}_j^{whi} + \mu_{train}. \tag{5}$$

These operations help the model process source and target CT scans on a consistent statistical basis. The model develops robustness to shifted distributions.
Classification. The last layer of the fine-tuned ResNet processes the aligned feature vector $\mathbf{x}_j^{whico}$ to predict a label for the target image I_j. The algorithm computes new summary statistics μ_{test} and $\mathbf{\Sigma}_{test}$ when it processes the next batch, while μ_{train} and $\mathbf{\Sigma}_{train}$ remain the same for all test batches.

3 Experiments and Results

We evaluate TTT-SparK and WhiCo on three 2D CT image classification tasks that cover diverse medical use cases and dataset sizes. The first task is an 11-class organ classification using the OrganSMNIST dataset [31]. The second task is a binary classification of lung CT scans to diagnose COVID-19 [33]. The third task requires binary classification to detect the presence/absence of a brain hemorrhage in CT scans. These scans come from an internal study conducted at *** hospital. Senior radiologists selected CT slices from 200 patients, resulting in 145 training, 25 validation, and 29 test images. The Ethics Committee of *** granted ethical approval. For each task, we have the uncorrupted test set without a distribution shift. Furthermore, we induce distribution shifts by applying image corruptions that are realistic for CT scans (i.e., inject speckle noise, increase (+) or decrease (-) gamma correction, contrast, and brightness) [7]. For each corruption, we apply the first of five increasing severity levels. Additionally, we simulate low-dose X-ray CT scans as described by Zeng et al. [36].
As a baseline, we use a ResNet-50 pre-trained with SparK [27] by Wolf et al. [30] and fine-tune it on the training and validation images of the OrgMNIST, COVID-19, or Brain dataset, following the procedure of Wolf et al. [30]. The baseline classifies test images without inference-time adaptation. We then analyze whether TTT-SparK or WhiCo improves the baseline performance under distribution shifts. For TTT-SparK, we use the LAMB optimizer (β: 0.9, 0.95) with a learning rate of 8e-5, weight decay of 1e-2, and train for 10 epochs (parameters found

Table 1. OrgMNIST Results (in %, average over 5 runs). The best result is marked in bold. WhiCo improves accuracy and F1 score across 50% of the corruption types.

	Accuracy ↑			AUC ↑			F1 ↑		
Corruption	Baseline	TTT	WhiCo	Baseline	TTT	WhiCo	Baseline	TTT	WhiCo
Uncorrupted	**81.9**	79.0	81.2	**98.0**	97.5	96.9	**81.9**	79.0	81.2
Low dose	41.3	40.2	**63.2**	89.8	88.4	**92.0**	41.3	40.2	**63.2**
Gamma+	77.8	73.3	**78.1**	**97.3**	96.4	96.1	77.8	73.3	**78.1**
Gamma–	**81.6**	79.2	80.9	**97.9**	97.4	96.9	**81.6**	79.2	80.9
Contrast+	**77.9**	74.1	77.6	**97.3**	96.5	96.1	**77.9**	74.1	77.6
Contrast–	77.7	74.1	**78.5**	**97.1**	96.4	96.3	77.7	74.1	**78.5**
Brightness+	**76.7**	74.3	75.8	**96.8**	96.1	95.7	**76.7**	74.3	75.8
Brightness–	67.0	60.5	**69.7**	**94.3**	92.6	93.6	67.0	60.5	**69.7**
Speckle Noise	**81.0**	78.4	80.6	**97.8**	97.3	96.8	**81.0**	78.4	80.6

through hyperparameter sweeps). WhiCo requires only the batch size to be specified, which varies by task as stated below. All experiments run on NVIDIA H100 and A100 GPUs. We repeat each experiment five times to compute the mean and standard deviation (std) of accuracy, AUC, and F1 scores on the test images.

OrgMNIST. Table 1 depicts the results on 8827 test images of the OrgMNIST task. TTT-SparK and WhiCo use a batch size of 128. All models exhibit stable results across all evaluation metrics. The results of the baseline, TTT-SparK, and WhiCo have a std of 0.02, 0.24, and 0.02 on average, respectively. TTT-SparK worsens the baseline's accuracy on both uncorrupted and corrupted images. In contrast, WhiCo enhances the baseline's accuracy on the gamma+, contrast-, brightness-, and low-dose corruption by 0.3, 0.8, 2.7, and 21.9 percentage points, respectively. Especially on low-dose X-ray CT scans, where the baseline accuracy drops drastically, WhiCo restores performance to above chance level (63.2%). WhiCo also achieves a 2.2-point higher AUC score than the baseline on the low-dose scans. Apart from this corruption, WhiCo does not substantially improve the AUC scores. TTT-SparK achieves slightly worse AUC scores than the baseline on all test sets. The F1 scores show the same trends as the accuracy results. Ranking our methods on organ identification reveals that WhiCo outperforms TTT-SparK in adapting the model to shifted distributions. Although WhiCo does not improve AUC scores remarkably, it enhances the model's performance across 50% of the corruption types in terms of accuracy and F1 score.

COVID-19. As in the first task, we run TTT-SparK and WhiCo with a batch size of 128. Table 2 shows the models' performance on 203 test images. The baseline performance decreases on the corrupted CT scans across all metrics. TTT-SparK achieves slightly higher accuracy than the baseline for the low-dose, gamma+, contrast+, brightness+, and speckle noise corruption. Nevertheless, there are no definite improvements as the baseline result falls within the std. On average, TTT-SparK's results show a std of 1.91. The baseline results have a std of 0.09 on average. WhiCo generates the most stable results, i.e., the std is 0.05 on average. Regarding the accuracy, WhiCo outperforms the baseline on the contrast-, speckle noise, brightness-, contrast+, and gamma+ corruption by 3.5,

Table 2. COVID-19 Results (in %, average over 5 runs). The best result is in bold. WhiCo improves accuracy, although it shows deficits in terms of AUC and F1 scores.

	Accuracy ↑			AUC ↑			F1 ↑		
Corruption	Baseline	TTT	WhiCo	Baseline	TTT	WhiCo	Baseline	TTT	WhiCo
Uncorrupted	**78.3**	78.0	67.5	81.5	**82.1**	71.0	**79.6**	79.2	68.9
Low dose	71.1	**71.8**	64.9	76.1	**79.6**	69.7	74.1	**74.4**	66.9
Gamma+	59.6	60.4	**68.0**	77.0	**78.1**	69.4	70.1	70.9	**71.1**
Gamma–	**68.0**	65.0	67.5	**79.5**	78.1	71.1	**74.5**	72.5	70.3
Contrast+	60.6	61.8	**68.5**	72.1	**73.2**	70.2	68.8	69.3	**69.8**
Contrast–	64.0	63.9	**67.5**	79.4	**80.4**	71.5	72.0	**72.6**	70.0
Brightness+	65.5	**65.8**	65.0	70.0	**71.4**	69.2	66.7	**67.2**	66.7
Brightness–	55.2	54.7	**60.6**	**68.8**	67.0	64.7	**68.9**	68.7	63.3
Speckle Noise	61.9	64.1	**66.9**	74.8	**76.9**	69.9	69.8	**71.3**	69.3

5.0, 5.4, 7.9, and 8.4 percentage points, respectively. Analyzing the AUC scores reveals that TTT-SparK fails to substantially outperform the baseline. Even where a slight improvement exists, the baseline score remains within the std. The only exception is on the low-dose images, where TTT-SparK improves the baseline by 3.5 points. WhiCo downgrades the AUC scores. Moreover, WhiCo and TTT-SparK do not drastically increase F1 scores. In summary, TTT-SparK does not enhance the model's performance in COVID-19 detection under distribution shifts. WhiCo succeeds in enhancing the model's accuracy, although it exhibits deficits in terms of AUC and F1 scores.

Brain. The brain test set contains only 29 images. Hence, Table 3 presents the results of TTT-SparK and WhiCo using a batch size of 29. Comparing the algorithms shows that WhiCo outperforms TTT-SparK, as it considerably improves accuracy, increases AUC scores, and exhibits a low std. The baseline performance drops drastically on the corrupted images. The accuracy and AUC scores downgrade to chance level. TTT-SparK fails to recover the model's accuracy, as it does not achieve substantial improvements on the corruptions. Additionally, it harms the performance on the uncorrupted images. Moreover, TTT-SparK does not achieve notably higher AUC or F1 scores than its baseline.

Table 3. Brain Results (in %, average over 5 runs). Best result is marked in bold. WhiCo outperforms TTT-SparK, as it considerably improves accuracy and AUC scores.

	Accuracy ↑			AUC ↑			F1 ↑		
Corruption	Baseline	TTT	WhiCo	Baseline	TTT	WhiCo	Baseline	TTT	WhiCo
Uncorrupted	**89.7**	75.2	82.8	**91.9**	91.8	91.4	**89.7**	68.9	83.9
Low dose	**51.7**	48.3	**51.7**	55.9	**56.0**	52.6	**66.7**	64.1	59.7
Gamma+	48.3	49.7	**55.2**	**55.7**	55.6	**55.7**	65.1	**65.4**	60.6
Gamma–	48.3	48.3	**58.6**	51.9	50.6	**52.9**	**65.1**	**65.1**	64.7
Contrast+	**48.3**	47.6	**48.3**	48.6	49.6	**51.0**	**65.1**	64.5	57.1
Contrast–	48.3	48.3	**55.2**	**59.1**	57.6	55.2	**65.1**	**65.1**	62.9
Brightness+	**48.3**	**48.3**	**48.3**	47.1	47.6	**52.4**	**65.1**	**65.1**	57.1
Brightness–	48.3	47.6	**51.7**	51.9	**55.7**	53.8	**65.1**	62.1	58.8
Speckle Noise	48.3	49.7	**58.6**	51.0	51.7	**53.7**	**65.1**	**65.1**	64.7

In contrast, WhiCo considerably improves accuracy by 3.4, 6.9, 6.9, 10.3, and 10.3 percentage points for the brightness-, gamma+, contrast-, gamma-, and speckle noise corruptions, respectively. For the remaining corruptions, it does not harm the accuracy. Although WhiCo does not outperform the baseline in terms of F1 scores, it increases the AUC scores by 1.0, 1.9, 2.4, 2.7, and 5.3 percentage points for the gamma-, brightness-, contrast+, speckle noise, and brightness+ corruptions, respectively. Furthermore, WhiCo leads to more stable results than TTT-SparK. On average, the baseline, TTT-SparK, and WhiCo results have a std of 0.05, 2.8, and 0.08, respectively.

4 Discussion and Conclusion

Our experiments evaluate TTT-SparK and WhiCo across diverse medical tasks (brain hemorrhage, COVID-19, and organ classification) and several corruptions, revealing valuable insights. Firstly, the baseline generalizes poorly to corrupted images, which emphasizes the need for a solution to the distribution shift problem. Secondly, TTT-SparK fails to meaningfully improve performance, as it does not outperform the baseline on all tasks. In contrast to TTT-SparK, WhiCo eliminates the need to optimize model weights and refine the learned decision boundary with a suitable loss at inference time. Instead, it aligns the target features with the source feature distribution. This strategy has some deficits in reducing false positives and false negatives, as WhiCo does not consistently increase the AUC and F1 scores across all experiments. A key strength of WhiCo is that it considerably improves the accuracy for most corruptions across all our tasks. Especially for large performance drops, it substantially increases the accuracy (e.g., on corrupted brain images). The small brain dataset assesses the methods' applicability to use cases where available CT images are limited due to disease rarity. WhiCo is effective in this low-data regime. Additionally, the low std reflects stable, reproducible performance, which is important for clinical deployment where reliability is crucial. Moreover, WhiCo can be transferred effortlessly to other convolutional models because it is model-agnostic and requires no domain knowledge to select hyperparameters. While existing literature has primarily used whitening and coloring transformations in UDA methods [21,24], we show that they can also be effectively leveraged for Test-Time Adaptation. While these findings are encouraging, several limitations should be noted. We use the lowest of five corruption severity levels and only simulate distribution shifts closely matching real-world CT shifts [7,36]. The simulations facilitate an evaluation in a controlled setting with known shifts. For future work, we plan experiments with higher severity levels and real, non-simulated distribution shifts.

In conclusion, we investigate Test-Time Adaptation methods to improve the robustness of a ResNet-50 to distribution shifts in CT imaging. TTT-SparK relies on masked autoencoding, whereas WhiCo leverages whitening and coloring as an adaptation strategy. Our results show that WhiCo achieves better outcomes than TTT-SparK, which highlights that aligning features is a valuable alternative to optimizing weights for adapting a ResNet-50 to a shifted data distribution.

References

1. Abbasi-Sureshjani, et al.: Risk of training diagnostic algorithms on data with demographic bias. In: International Workshop on Interpretability of Machine Intelligence in Medical Image Computing. pp. 183–192. Springer (2020)
2. Abdullah, M., et al.: A Hybrid Deep Learning CNN model for COVID-19 detection from chest X-rays. Heliyon **10**(5) (2024)
3. Alshazly, H., et al.: COVID-Nets: deep CNN architectures for detecting COVID-19 using chest CT scans. PeerJ Computer Science **7**, e655 (2021)
4. Armato III, S.G., et al.: The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Medical physics **38**(2), 915–931 (2011)
5. Chen, L., et al.: Reusing the task-specific classifier as a discriminator: Discriminator-free adversarial domain adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7181–7190 (2022)
6. Chougrad, H., et al.: Deep convolutional neural networks for breast cancer screening. Computer methods and programs in biomedicine **157**, 19–30 (2018)
7. Di Salvo, F., et al.: Medmnist-c: Comprehensive benchmark and improved classifier robustness by simulating realistic image corruptions. arXiv preprint arXiv:2406.17536 (2024)
8. Dosovitskiy, A., et al.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: International Conference on Learning Representations (2020)
9. Gandelsman, Y., et al.: Test-time training with masked autoencoders. Advances in Neural Information Processing Systems **35**, 29374–29385 (2022)
10. Goldblum, M., et al.: Battle of the backbones: A large-scale comparison of pretrained models across computer vision tasks. Advances in Neural Information Processing Systems **36**, 29343–29371 (2023)
11. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Humayun, M., et al.: A transfer learning approach with a convolutional neural network for the classification of lung carcinoma. Healthcare **10**(6), 1058 (2022)
13. Liu, J., et al.: Alzheimer's disease detection using depthwise separable convolutional neural networks. Computer Methods and Programs in Biomedicine **203**, 106032 (2021)
14. Liu, M., et al.: A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease. Neuroimage **208**, 116459 (2020)
15. Liu, Y., et al.: Ttt++: When does self-supervised test-time training fail or thrive? Advances in Neural Information Processing Systems **34**, 21808–21820 (2021)
16. Liu, Z., et al.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
17. Petersen, K.B., et al.: The matrix cookbook. Technical University of Denmark **7**(15), 510 (2008)
18. Pooch, E.H., et al.: Can we trust deep learning based diagnosis? the impact of domain shift in chest radiograph classification. In: International workshop on thoracic image analysis. pp. 74–83. Springer (2020)
19. Praveena, G., et al.: Early detection of Alzheimer's disease and dementia using deep convolutional neural networks. In: 2024 Third International Conference on

Distributed Computing and Electrical Circuits and Electronics (ICDCECE). pp. 1–5. IEEE (2024)
20. Recht, B., et al.: Do cifar-10 classifiers generalize to cifar-10? arXiv preprint arXiv:1806.00451 (2018)
21. Roy, S., et al.: Unsupervised domain adaptation using full-feature whitening and colouring. In: International Conference on Image Analysis and Processing. pp. 225–236. Springer (2019)
22. Shazia, A., et al.: A comparative study of multiple neural network for detection of COVID-19 on chest X-ray. EURASIP journal on advances in signal processing **2021**(1), 50 (2021)
23. Song, Y., et al.: Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images. IEEE/ACM transactions on computational biology and bioinformatics **18**(6), 2775–2780 (2021)
24. Sun, B., et al.: Correlation alignment for unsupervised domain adaptation. In: Domain adaptation in computer vision applications, pp. 153–171. Springer (2017)
25. Sun, Y., et al.: Unsupervised domain adaptation through self-supervision. arXiv preprint arXiv:1909.11825 (2019)
26. Sun, Y., et al.: Test-time training with self-supervision for generalization under distribution shifts. In: International conference on machine learning. pp. 9229–9248. PMLR (2020)
27. Tian, K., et al.: Designing BERT for Convolutional Networks: Sparse and Hierarchical Masked Modeling. In: The Eleventh International Conference on Learning Representations (2023)
28. Tzeng, E., et al.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7167–7176 (2017)
29. Wang, D., et al.: Tent: Fully test-time adaptation by entropy minimization. arXiv preprint arXiv:2006.10726 (2020)
30. Wolf, D., et al.: Self-supervised pre-training with contrastive and masked autoencoder methods for dealing with small datasets in deep learning for medical imaging. Scientific Reports **13**(1), 20260 (2023)
31. Yang, J., et al.: Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific data **10**(1), 41 (2023)
32. Yang, T., et al.: Test-time batch normalization. arXiv preprint arXiv:2205.10210 (2022)
33. Yang, X., et al.: Covid-ct-dataset: a ct scan dataset about covid-19. arXiv preprint arXiv:2003.13865 (2020)
34. Yue, X., et al.: Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13834–13844 (2021)
35. Zech, J.R., et al.: Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS medicine **15**(11), e1002683 (2018)
36. Zeng, D., et al.: A simple low-dose x-ray CT simulation from high-dose scan. IEEE transactions on nuclear science **62**(5), 2226–2233 (2015)
37. Zhang, M., et al.: Memo: Test time robustness via adaptation and augmentation. Advances in neural information processing systems **35**, 38629–38642 (2022)
38. Zhang, N., et al.: Skin cancer diagnosis based on optimized convolutional neural network. Artificial intelligence in medicine **102**, 101756 (2020)