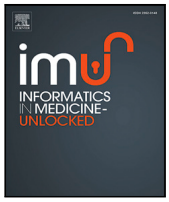
Informatics in Medicine Unlocked 44 (2024) 101430

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu



Mitigating the influence of domain shift in skin lesion classification: A benchmark study of unsupervised domain adaptation methods

Sireesha Chamarthi [a,1], Katharina Fogelberg [b,1], Titus J. Brinker [b,*,2], Julia Niebling [a,2]

[a] Data Analysis and Intelligence, German Aerospace Center (DLR - Institute of Data science), Jena, Germany
[b] Digital Biomarkers for Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany

ARTICLE INFO

Keywords:
Domain shift
Skin lesion classification
Dermoscopic images
Unsupervised domain adaptation
Generalization

ABSTRACT

The potential of deep neural networks in skin lesion classification has already been demonstrated to be on-par if not superior to the dermatologists' diagnosis in experimental settings. However, the performance of these models usually deteriorates in real-world scenarios, where the test data differs significantly from the training data (i.e. domain shift). This concerning limitation for models intended to be used in real-world skin lesion classification tasks poses a risk to patients. For example, different image acquisition systems or previously unseen anatomical sites on the patient can suffice to cause such domain shifts. Mitigating the negative effect of such shifts is therefore crucial, but developing effective methods to address domain shift has proven to be challenging. In this study, we carry out a comparative analysis of eight different unsupervised domain adaptation methods to analyze their effectiveness in improving generalization for dermoscopic datasets. To ensure robustness of our findings, we test each method on a total of ten derived datasets, thereby covering a variety of possible domain shifts. In addition, we investigated which factors in the domain shifted datasets have an impact on the effectiveness of domain adaptation methods. Our findings show that all of the eight domain adaptation methods result in improved AUPRC for the majority of analyzed datasets. Altogether, these results indicate that unsupervised domain adaptations generally lead to performance improvements for the binary melanoma-nevus classification task regardless of the nature of the domain shift. However, small or heavily imbalanced datasets lead to a reduced conformity of the results due to the influence of these factors on the methods' performance.

1. Introduction

Deep Neural Networks (DNNs) transformed machine learning by significantly improving predictive accuracy, even in complex experimental applications. Several recent works have demonstrated the applicability of deep learning based methods for skin lesion classification [1–3]. There are also efforts to develop different kinds of approaches to improve the performance of deep learning models for real world scenarios [4–7]. Usually, DNNs are trained on large datasets, so they learn the representations effectively. Apart from that, the training dataset (or source) and the test dataset (or target) for classification models are drawn from the same distribution. However, in skin cancer classification, as well as in other potential real-world scenarios, the source and target domains are generally different. Even a small-scale deviation from the distribution of the training domain can result in unreliable and deteriorated predictions on the target domain [8–11]. This deviation between datasets is commonly referred to as domain shift. In dermatology, these domain shifts can be caused by a combination of different factors, such as changes in the settings of an image acquisition system, view angle, patient age, lighting conditions in the examination room, or the way the dermatoscope is positioned, among others.

As such domain shifts result in a performance decrease, there exist different approaches to address this issue, e.g. data augmentation [12,13], domain generalization [14] and domain adaptation (DA) [15,16]. Domain generalization and DA are closely related. While domain generalization methods do not access any data from the target domain, domain adaptation methods may make use of data from the target domain by definition. Nevertheless, all these approaches can only reduce, but not remove the discrepancy between domains [8].

Domain adaptation is typically applied in cases where the domain feature spaces and tasks remain the same while only the distributions differ between source and target datasets (presence of a domain shift) [17–19]. Mainly this is done by either moment-matching methods

* Correspondence to: Division of Digital Biomarkers for Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany.
E-mail address: titus.brinker@nct-heidelberg.de (T.J. Brinker).
[1] Both authors contributed equally.
[2] Both authors contributed equally.

https://doi.org/10.1016/j.imu.2023.101430
Received 2 November 2023; Accepted 6 December 2023
Available online 7 December 2023

or by adversarial learning [20]. The knowledge transfer from source to target works via finding domain-invariant representations, which are used to bridge the discrepancy between domains [21].

Unsupervised Domain Adaptation (UDA) methods are well studied and established on multiple benchmark datasets (usually natural images), like Office-10, Caltech-10, Office-31, MNIST, and SVHN datasets [15,22,23], but their performance is not verified on new tasks [24]. Therefore it may be more difficult to choose a proper method for real-life applications. Apart from this, existing medical images are mostly unlabeled as it is generally difficult to obtain labeled data in the medical field. For a sufficient ground truth (labels) for dermoscopic images, a biopsy of the human lesion needs to be performed. Therefore, the overall process of obtaining and reliably labeling dermoscopic data is labor-intensive. That is why further task-specific fine-tuning of DNN is time-consuming and difficult. These limitations can be addressed by utilizing specifically domain adaptation methods which are unsupervised [25].

Significant work has already been invested into utilizing data augmentation techniques for dermoscopic skin lesion analyses [26–28]. Similarly, domain generalization techniques [29,30] and DA methods (Section 2) have been utilized for dermoscopic image analyses. However, to our knowledge, there is no previous research that applies UDA methods as a benchmark on dermoscopic skin cancer datasets. In addition, most existing works on domain adaptation assume their datasets to be domain shifted without quantifying it. In more complex tasks such as dermoscopic image classification, where even medical experts struggle to differentiate melanomas and nevi in particular situations, it is crucial to ensure that the datasets are truly domain shifted. In our previous work [31], we grouped and quantified domain shifted datasets for dermoscopic skin cancer classification, which we will use in this study. Additionally, other studies acknowledge their performance improvements without focusing on influential factors. We aim to identify possible factors for this performance improvement. Furthermore, other studies typically focus only on their benchmark and do not compare their results to other tasks, which can limit the generalizability of their findings. Therefore, while good performance on one method with one dataset or task may indicate its effectiveness, it does not guarantee the same performance improvement with other datasets or tasks.

Our contributions are the following:

- we provide a comparative analysis of 8 UDA methods on 10 derived dermoscopic datasets with quantified (not assumed) biological and technical domain shifts.
- we identify dataset- and method-specific factors that influence the performance of UDA methods.
- we compare our results to other benchmark domain adaptation datasets (e.g. Office-31).

This work is structured as follows: First, we discuss related works which focus on UDA methods and benchmarking in Section 2. In Section 3 we describe the used dermoscopic datasets and the UDA methods we compared in our analyses. We further explain our experimental settings. Finally, in Section 4 we discuss our results regarding different aspects of comparison. We examine the influence of class imbalance, target dataset size, as well as the performance itself using the Area Under the Receiver Operating Characteristic (AUROC) and the Area Under the Precision–Recall Curve (AUPRC). We conclude the paper in Section 5 with our findings and discuss possible future research directions.

2. Related work

Ben-David et al. [32] pioneered domain adaptation theory and further classified DA methods into supervised and unsupervised approaches based on label availability. In Supervised Domain Adaptation (SDA) the model is trained on the source domain and tested on the target domain, both with labeled data. The most common approach for SDA is pretraining on the source domain and fine-tuning on the target domain. However, for the translation of medical applications into the clinic this approach is impractical and time-consuming because it needs to be retrained for every new clinical scenario. The main goal of UDA is to enable the adaptation to new domains for better generalization by matching the marginal [33–36] or the conditional distributions [37,38] of the labeled source and unlabeled target domains. As the dearth of labeled data is the most prominent issue in the medical field, UDA methods gained a lot of attention, especially in medical image analysis [25]. Owing to the advantages of UDA- over SDA methods, most of the existing DA research is focused on UDA. To enable adaptation from the source domain to the target domain, UDA methods have to meet two important criteria, namely transferability and discriminability [39]. The transferability of feature representations from source to target is the primary indicator of the performance of the model. Apart from this, the other key indicator is the ability to discriminate between the classes present in the domains. There are mainly two strategies to align feature distributions across domains: Moment matching and adversarial training [24].

Moment matching methods aim to decrease the distribution discrepancy between the source and the target domain. This is achieved by matching the first-/ second-order moments (as mean and covariance) of the activation distributions that are unique to each domain in the hidden activation space [40]. Multiple UDA methods have been developed based on moment matching, including Deep Adaptation Networks (*DAN*) [18] which utilize Maximum Mean Discrepancy (*MMD*). An extension of *DAN*, called Joint Adaptation Networks (*JAN*) [41] has also been established. Apart from that, Correlation Alignment (*Deep-CORAL*) [42] is based on second-order statistics of both distributions. Another approach is *CMD* [40], which defines a new distance function based on probability distributions by moment sequences. Methods based on divergence are typically not very complicated, easier to train, and do not require a lot of hyperparameter tuning for optimization. Additionally, they are computationally efficient and are not in necessity of large datasets [43]. However, the disadvantage of these types of methods is that they cannot be reliably used to achieve good performances on large datasets with more complex and diverse images. Also, they cannot be applied to other computer vision tasks, such as semantic segmentation, because they do learn image-level, not pixel-level representations.

Adversarial training methods for domain adaptation learn domain-invariant features. For this, a domain discriminator is trained to differentiate between the source and the target domain by minimizing the classification error. At the same time, the feature representations learned by the network try to confuse the discriminator. One of the well-studied and most used adversarial methods is the Domain Adversarial Neural Network (*DANN*) [19]. Apart from *DANN*, there are other adversarial methods like Adversarial Discriminative Domain Adaptation (*ADDA*) [10] and Maximum Classifier Discrepancy (*MCD*) [44] which are developed as an extension to the *DANN* approach. Typically, adversarial methods achieve better adaptations than moment matching methods and are the more dominant method [11]. They are very good at enhancing the transferability of representations. Additionally, they have good computational efficiency and work across different kinds of datasets [43]. Discriminative approaches are able to adapt well to larger domain shifts [10]. A disadvantage is, that in some cases they may perform poorly on small datasets because these methods rely on the convergence of a min–max game. Furthermore, it can be difficult to optimize these models, and when having multimodal feature distributions it can be challenging for adversarial methods to adapt feature representations only [20,45,46]. The discriminability of the learned representation happens only by minimizing the classification error on the source domain [39]. It cannot be guaranteed that the distributions are identical, even if the confusion of the discriminator was fully achieved [20,47].

Table 1
Overview of the derived datasets used for benchmarking, including dataset sizes and class distributions. H represents our source dataset. All following domain shifted datasets are adapted with respect to H.

Abbreviation	Origin	Biological factors	Melanoma amount	Nevus amount	Total target size
H	HAM	Age > 30, Loc. = Body (default)	465 (10%)	4234 (90%)	4699
HA	HAM	Age ≤ 30, Loc. = Body	25 (4%)	532 (96%)	557
HLH	HAM	Age > 30, Loc. = Head/Neck	99 (45%)	121 (55%)	220
HLP	HAM	Age > 30, Loc. = Palms/Soles	15 (7%)	203 (93%)	218
B	BCN	Age > 30, Loc. = Body (default)	1918 (41%)	2721 (59%)	4639
BA	BCN	Age ≤ 30, Loc. = Body	71 (8%)	808 (92%)	879
BLH	BCN	Age > 30, Loc. = Head/Neck	612 (66%)	320 (34%)	932
BLP	BCN	Age > 30, Loc. = Palms/Soles	192 (65%)	105 (35%)	297
M	MSK	Age > 30, Loc. = Body (default)	565 (31%)	1282 (69%)	1847
MA	MSK	Age ≤ 30, Loc. = Body	37 (8%)	427 (92%)	464
MLH	MSK	Age > 30, Loc. = Head/Neck	175 (60%)	117 (40%)	292

There exist also **extensions to adversarial methods**, e.g. Batch Spectral Penalization (*BSP*) [39], which can be used standalone or as a regularizer to another domain adaptation method. Also Minimum Class Confusion (*MCC*) [24] can be used as a standalone adaptation method or additionally as a regularizer. The advantage of *MCC* over *BSP* is, that *MCC* as a regularizer is not limited to adversarial methods.

Besides, there is also extensive research in the direction of **adversarial generative methods** which are based on GANs. They include a generator to create virtual images, while a discriminator tries to differentiate between real and generated images [45]. The research in the area of conditional GANs [48] led to the development of methods like Conditional Adversarial Domain Adaptation (*CDAN*) [20]. Although adversarial generative adaptation methods usually achieve good performances, they require largely scaled data for the generator to be trained properly. Furthermore, these methods need more computational resources, as well as hyperparameter tuning, which makes the optimization process more complex [43]. Additionally, GANs show attractive visualizations, but they can be limited to small shifts [10].

Due to the growing demand for adapting neural networks to unseen domains, there are other popular methods like Unsupervised Image-to-Image Translation Networks (*UNIT*) [49], Generate to Adapt (*GTA*) [50], Cycle-Consistent Adversarial Domain Adaptation (*CyCADA*) [51] and Adaptive Feature Norm (*AFN*) [11].

It is important to note that this area of research is rapidly growing and new domain adaptation methods are emerging in a variety of fields, ranging from computer vision, natural language processing, and video analysis to robotics. Their use-case is also not just limited to image classification tasks but is extended to semantic segmentation, face recognition, object identification, image-to-image translation, person re-identification, and image captioning, among others [15]. Domain adaptation is also commonly used in medical image analysis. The leading application area of visual domain adaptation in medicine are brain images [25], while there is also research on lungs, hearts, breasts, eyes, and abdomen. Mostly, these applications use histological or microscopical images.

We have noticed that there is limited work applying DA to dermoscopic images. Gu et al. [52] developed a two-step progressive adaptation method for task specific skin cancer classification. In their approach, they first trained a CNN on ImageNet and further fine-tuned it on an intermediate skin cancer dataset, before fine-tuning it again on another skin cancer dataset. Apart from that, Ahn et al. [53] used a similar approach of training the model initially on ImageNet and fine-tuning it on medical images. They used context-based feature augmentation which uses additional information about the images. They experimented with medical image modality classification, a tuberculosis dataset, as well as with skin cancer datasets.

UDA methods are typically compared against each other when a new method is proposed. In that comparison, the works mostly focus on performance comparisons with respect to other state-of-the-art methods. Most of the UDA methods are evaluated on well-studied datasets like ImageNet, MNIST, and Office-31, whereas their performance on other datasets is expected to change based on the available data and the domains present in them. Even these benchmark datasets are not analyzed for artifacts and duplicates present within the dataset. Ringwald et al. [54] analyzed frequently used UDA datasets and studied the systematic problems with regard to dataset setup and ambiguities. They established a clean Office-31 dataset for UDA algorithm comparisons. To verify the actual efficiency of the UDA methods, it is essential to study their performance on other, more real-world related datasets, as well. Peng et al. [55] introduced a benchmark dataset to evaluate the performance of UDA methods. They estimate the performance of the domain adaptation models to transfer knowledge from synthetic to real data. Also, Nagananda et al. [56] compared UDA methods on publicly available aerial datasets. In the medical field, Saat et al. [57] proposed a benchmark for UDA methods on brain Magnetic Resonance Imaging (MRI) - an image segmentation task. In their work, they compared UDA methods and evaluated the performance with respect to their baseline model. The source domain consists of MRI scans from multiple centers and different scanners. Whereas the target domain consists of MRI scans from a different dataset from a single center. We noticed that there is no extensive work on benchmarking UDA methods in particular for dermoscopic image classification.

3. Materials and methods

3.1. Datasets

Even though some recent works used image datasets of skin lesions, like MoleMap, HAM10k, and ISIC [25] for their adaptation tasks, there is no study evaluating the actual and total domains present in these datasets or developing and evaluating public dermoscopic datasets particularly for domain adaptation techniques [57]. To overcome this limitation, we grouped and quantified potential technical and biological shifts in our previous work [31] to obtain domain shifted dermoscopic datasets.[3] Table 1 provides a summary of the domains observed in the dermoscopic datasets.

As we are using unsupervised approaches, the source domain is labeled and these labels are used for the classification at the end. It is essential to have a dataset that can be divided into train and test without data leakage, which is not always straightforward for dermatology datasets due to duplicated lesion images. Apart from that, in domain adaptation analyses, the methods are evaluated from one domain to another (domain A to domain B) and are also tested in the opposite direction (domain B to domain A) [19]. However, recent works stated that the performance of UDA methods is negatively affected by poor data quality and duplicates in the datasets [54]. This can be a difficulty when using the publicly available ISIC archive images as they contain duplicates that are not necessarily marked as such [58].

[3] https://gitlab.com/dlr-dw/isic_download

Table 2
Our selection of eight state-of-the-art UDA methods for the benchmark study. The methods are based on different types of approaches.

UDA method	Type
DAN [19]	Moment matching
JAN [41]	Moment matching
DANN [19]	Adversarial training
ADDA [10]	Adversarial training
BSP [39]	Extension of adversarial
MCC [24]	Extension of adversarial
CDAN [20]	Adversarial generative
AFN [11]	Other

We chose a representative and large subset of HAM, dataset *H* (Table 1), as our only source domain for the adaptation process. For this, we used the lesion IDs present in HAM10k to remove the duplicates in the dataset [59]. Therefore, the adaptation was done in one direction only, using derived sub-datasets *HA*, *HLH*, *HLP*, as well as BCN20k [60] and MSK [58] datasets exclusively as target domains.

3.2. UDA methods

Overall, we focus on single-source, single-target, homogeneous adaptation without target labels. This means that there is one fully labeled source domain and one unlabeled target domain within the same modality and that the source and target domains share the same classes.

We selected eight state-of-the-art UDA methods (Table 2), which were selected based on different types, computational efficiency, and good performance on different established datasets. These methods have been extensively used in both medical and non-medical applications as such or as the basis of newer approaches [61]. However, the field is evolving rapidly and there are many new techniques outperforming others in various applications. It would be beyond the scope of this work to compare more methods.

3.3. Experimental setup

It is difficult to decide which UDA method is generally better compared to others in terms of design or performance. The key characteristic that determines the strength of a UDA method is its ability to transfer feature representations from a source- to a target domain. For this reason, we compare all UDA results to our unadapted baseline method (*Src*) trained on source dataset *H*, which is a basic ResNet50 model [62] pre-trained on ImageNet. The other performance characteristic is the discriminability between the classes within domains. We evaluate how well the model is able to discriminate between melanoma and nevus in our binary classification task. For this we follow standard evaluation protocols for unsupervised domain adaptation [19,41]. For all experiments, we used an initial learning rate (LR) of 0.01 with a weight decay of 1e−3 and a LR-decay of 0.75. The used momentum was 0.9 and gamma was 0.001. We set the epochs to 20 and the batch size to 16. The comparison is based on an existing repository[4] which already implemented a variety of methods. It is open-source and has been established on multiple popular datasets, e.g. MNIST, Office-31, and DomainNet [61,63]. We modified the library for our classification of dermoscopic images.

In a typical dermoscopic dataset, the presence of melanoma, in comparison to nevus images, is very low, as can be seen in Table 1. In our analysis, we consider melanoma as the positive and nevus as the negative class. When it comes to a clinical translation of a diagnostic system for skin lesions, both, True Positives and True Negatives are considered very important. Therefore, we focused on AUROC and AUPRC as evaluation metrics. The advantage of these two metrics is that they are both threshold-free. That means, that they can provide an overview of the performance range with different dataset splits into positively and negatively predicted classes [64]. Also, Zhang et al. [65] used AUROC and AUPRC for the evaluation of their domain adaptation results in a recent work.

AUROC as a standalone metric can be misleading in imbalanced tasks because the score can be better than random guessing (baseline = 0.5), but still misclassify the minority class. On the contrary, AUPRC is tailored for such imbalanced cases, but may mislead in balanced cases or where the negatives are rare. When using only AUPRC, it can be difficult to compare results across datasets with different class ratios and dataset sizes. The reason for this is the varying baseline of this metric, as it is dependent on the ratio of the positive class. Therefore, we computed both metrics, while also focusing on the AUPRC improvement (in %) compared to the unadapted baseline method (*Src*), as suggested by Zhang et al. [66]. With this approach, the results can be compared across methods and datasets equally.

For the experiments, we included a weighted random sampler to maintain equal class ratios per batch during model training. We also adopted five-fold cross-validation to use all images of the available datasets. From each fold, we selected the best epoch (out of 20) and averaged the results. Additionally, we ran the experiments with five seeds to observe the variability of the results over different runs. For the end results, we averaged the values over five seeds. The seeding makes the performance results more robust and that way shows more realistic values.

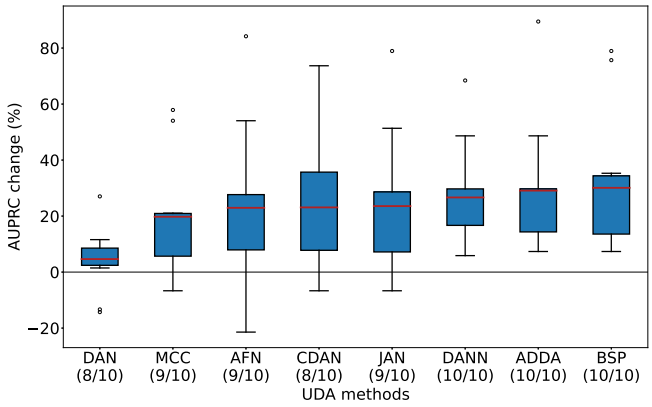


Fig. 1. Comparison of UDA methods with respect to AUPRC change. The red line represents the mean (of the performance on all derived datasets) and the black dots are outliers. The black line shows the baseline at 0% performance improvement. The performance improvement was calculated over five seeds and averaged over ten datasets. The UDA methods are ordered in the increasing order of the mean AUPRC change on the *x*-axis. The numbers in the brackets (x-axis) represent how many datasets out of ten this particular method improved the performance.

4. Results and discussion

If a UDA method performs well on one dataset, it does not guarantee similar performance on other datasets. As discussed in Section 3, we have selected eight state-of-the-art UDA methods to evaluate the performance on the domains present in our ten dermoscopic datasets. We compared all adaptation methods with our non-adapted baseline approach.

4.1. Benchmarking UDA methods on dermoscopic datasets

As stated in Section 3.3, we computed AUROC and AUPRC scores for different derived datasets and methods, which can be seen in

[4] https://github.com/thuml/Transfer-Learning-Library

Table 3
Comparison of AUPRC results across different derived datasets and UDA methods. The columns represent the domain shifted datasets (target) for the source dataset *H* (not listed here). Each row represents the results for a particular UDA method, with the first row indicating the results for the unadapted baseline method (*Src*). The best-performing UDA method for each dataset is highlighted in bold. The percentage for each dataset shows the ratio of melanoma in that dataset, which serves as the baseline for AUPRC. The source dataset *H* comprises only 10% melanoma.

		Domain shifted dataset									
		HA	HLH	HLP	B	BA	BLH	BLP	M	MA	MLH
	Mel (%)	4	45	7	41	8	66	65	31	8	60
(UDA) method	Src	0.14±0.02	0.69±0.04	0.37±0.15	0.57±0.02	0.19±0.06	0.73±0.03	0.77±0.05	0.34±0.01	0.15±0.04	0.68±0.03
	DAN	0.12±0.02	0.77±0.02	0.47±0.14	0.60±0.04	0.20±0.02	0.78±0.01	0.83±0.03	0.37±0.03	0.13±0.03	0.69±0.03
	JAN	0.15±0.04	0.82±0.05	0.56±0.08	0.72±0.02	0.34±0.02	0.85±0.02	0.82±0.03	0.44±0.03	0.14±0.01	**0.73±0.03**
	DANN	0.17±0.01	0.81±0.04	0.55±0.07	0.74±0.02	0.32±0.03	0.85±0.01	0.84±0.01	0.44±0.01	**0.18±0.04**	0.72±0.02
	ADDA	**0.18±0.06**	0.81±0.02	0.55±0.03	0.74±0.01	**0.36±0.03**	**0.87±0.01**	0.83±0.02	0.44±0.02	0.17±0.03	**0.73±0.03**
	CDAN	0.14±0.02	0.82±0.03	0.54±0.06	0.73±0.02	0.33±0.02	0.85±0.01	0.84±0.02	**0.47±0.02**	0.14±0.01	**0.73±0.02**
	BSP	0.16±0.03	0.82±0.02	**0.65±0.04**	**0.75±0.01**	0.34±0.05	0.86±0.01	0.83±0.02	0.46±0.02	0.17±0.03	**0.73±0.01**
	AFN	0.11±0.02	**0.83±0.02**	0.57±0.13	0.73±0.01	0.35±0.03	0.84±0.01	**0.86±0.02**	0.43±0.02	0.16±0.02	0.71±0.02
	MCC	0.15±0.04	**0.83±0.07**	0.57±0.05	0.69±0.02	0.30±0.08	0.83±0.01	0.81±0.04	0.41±0.02	0.14±0.03	0.71±0.03

Table A.1 and Table 3, respectively. To provide a better understanding of our comparison and to demonstrate the quantified changes compared to *Src* method, we also looked at the AUROC- (Table A.2) and AUPRC (Table A.3) improvement (in %). In these tables, negative values, which indicate performance degradation after domain adaptation, occur rarely.

Our results indicate, that all selected UDA methods achieve a performance improvement (in %) compared to *Src* method over most available domain shifted datasets (Fig. 1). *BSP*, *ADDA* and *DANN*, which are all adversarial types of techniques, achieve the largest performance improvement. According to our results, these three UDA methods were able to improve the performance of 10 out of 10 domain shifted dermoscopic datasets.

The performance change (in %) of each individual domain shifted dataset per method is represented in Fig. 2. For this overview, we combine performance change, melanoma ratio, and target dataset size in one figure. While the performance is demonstrated in the upper point plot, the melanoma ratio per dataset can be observed in the lower illustration. In both sub-figures the domain shifted datasets on the *x*-axis are ordered by target dataset size in an ascending order from left to right. The largest improvements are achieved on dataset *BA* using either *ADDA* or *AFN* as the UDA method. However, although *BA* has the highest improvement, it has also a high variance between the methods' results, which ranges from 5.26% to 89.47%. All UDA methods, except for *DAN*, achieved maximum performance improvement at least for one dataset, as shown in Table A.3. It is also noteworthy that the *MLH* dataset posed the greatest challenge for adaptation, as all UDA methods seem to encounter difficulties with it (Fig. 2).

4.2. Influential factors on the performance improvement

We evaluated the results from multiple perspectives, including dataset- and method-specific factors, which could influence the performance. Of course, these factors are not exclusive and may have a different strength of influence in other applications or tasks.

4.2.1. Dataset-specific factors

Our analysis revealed that the amount of melanoma images in the target datasets affects the performance of UDA methods, as demonstrated in Table 3 and Fig. 2. Several derived datasets, including *HA*, *HLP*, *MA*, and *BA*, have a low number of melanoma cases and also represent larger disparities between the results of all UDA methods (Table A.3). Datasets *HLH*, *B* and *M* have a more balanced distribution between both classes and therefore show more agreeing results between all UDA-methods (Table 1). Adversarial methods and their extensions appear to perform better for such imbalanced datasets. For instance, *ADDA* is the most effective UDA method for dataset *HA*, which has the lowest melanoma ratio of 4%. On the other hand, some datasets such as *BLP*, *BLH*, and *MLH* are dominated by melanoma cases and therefore all methods show similar improvement in AUPRC scores. Although there are cases where no improvement can be detected from the unadapted baseline, we observe that most methods agree with each other when it comes to datasets with a high melanoma ratio (Fig. 2).

It is worth noting that *MA*, *HA*, and *BA* datasets contain images of skin lesions from patients below the age of 30. These datasets include both, young patients and children, as we have previously noted in our work [31]. Diagnosing melanoma in children is a unique challenge in clinical diagnosis, as they do not show typical ABCDE features [67] used to identify melanomas in adults due to their different appearance [68]. This may result in a lower performance improvement after adaptation.

In order to achieve performance improvements in UDA methods, it is necessary to have a large dataset available for the training process of the adaptation method [69]. As shown in Table 3 and Fig. 2 this is a fact we can confirm, because for the larger datasets *M* and *B*, most of the methods (except for *DAN*) showed higher improvement in performance compared to other datasets. Interestingly, these two datasets have a balanced class distribution, too, which is most likely influential, as well. An exception to this observation is dataset *HLP* where most methods show agreement despite the small dataset size and low melanoma ratio. We assume this is because of the relative similarity of the target dataset to the source dataset (*H*). In our previous analysis [31], we found that *HLP* is one of the most similar datasets to *H* in terms of melanoma images, as measured by cosine similarity and JS-divergence. Additionally, it is worth pointing out that for this dataset, the variation between the least performing *DAN* and the best performing *BSP* method is high.

In our study, we were under the assumption that higher divergence corresponds to lower performance, therefore leading to an important investigation into the relationship between domain shift measures and performance. For JS-divergence and AUROC (mean across all methods), as well as JS-divergence and AUROC (best-performing method) we observed only a moderate correlation. Interestingly, when observing the behavior of the two separate classes, melanoma and nevus, the correlation was slightly stronger for the nevus class. When comparing these findings to the domain shift measure cosine similarity, we found that only the nevus class exhibited moderate correlations. Specifically, the AUPRC (mean across all methods) and the AUPRC (best-performing method) showed moderate correlations with cosine similarity. The improvement percentage and the variation intensity between the performances did not show any correlation with one of the domain shift quantification measures. Nevertheless, when examining the domain shift qualitatively, it appears to be challenging to improve performance using UDA methods when both, biological and technical shifts are represented.

In summary, various dataset-specific factors contribute to the performance of UDA methods, such as melanoma ratio, target size, and

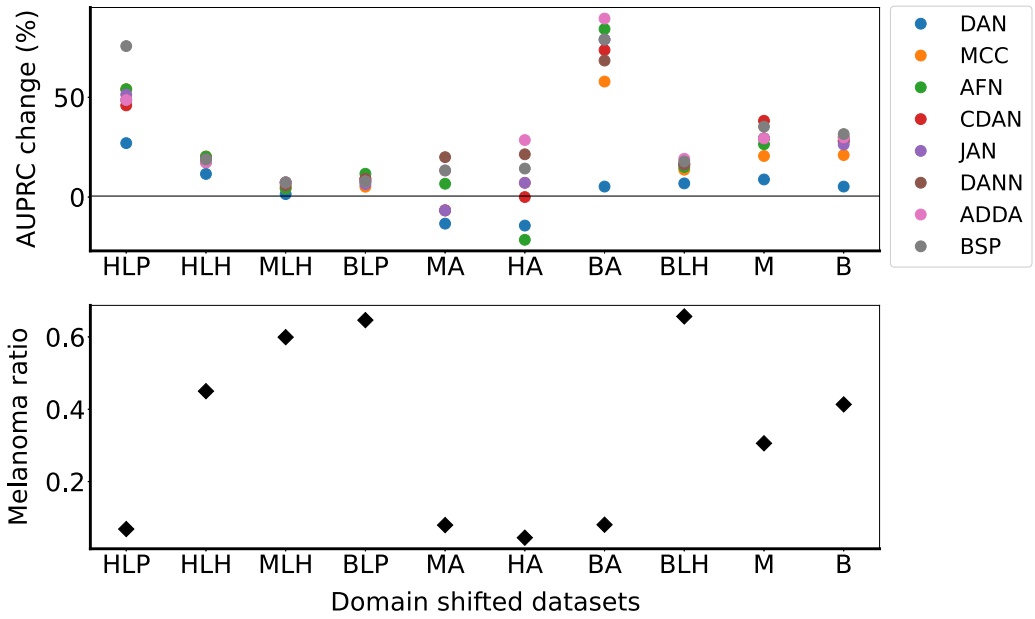


Fig. 2. Change in AUPRC (in %) with respect to the unadapted baseline model (*Src*). Individual UDA methods (color-coded) are illustrated across all domain shifted datasets (x-axis). The upper panel of the figure shows the AUPRC change (in %). The black line at 0 on the *y*-axis highlights the methods that show a performance degradation or no improvement (w.r.t the unadapted baseline method). The lower panel shows the melanoma ratio for each dataset. The datasets on the *x*-axis are ordered by total target size in an ascending manner.

how similar or dissimilar datasets are with respect to the source dataset *H*. At present, it remains unclear which factor has the greatest influence on performance improvement with UDA. More likely, the performance is influenced by a combination of multiple factors. Therefore, it is essential to continue investigating and exploring influential factors to better understand their impact on UDA performance. Our analysis was aimed to understand if UDA methods are effective for adapting to unseen skin lesion domains. In this process, we tried to estimate what factors might have resulted in the final performance seen in Fig. 2, however, we believe there are other influencing variables in addition to what was discussed.

4.2.2. Method-specific factors

As shown in Fig. 1, *BSP* is the best-performing method with higher mean AUPRC change compared to all the approaches used. This can be attributed to the fact that *BSP* utilizes regularization, which can be incorporated into adversarial domain adaptation networks [39]. It penalizes the spectral norm of the adaptation layers, which is believed to have an effect on the domain shifts specific to our datasets.

The next best performing method is *ADDA*, which is also an adversarial approach focusing on discriminative adaptation. This method combines the key strategies of previously demonstrated domain adaptation approaches into one method [10]. The authors of the approach developed this method as a more generalized framework for adversarial adaptation that includes other adversarial approaches. We assume that might be one of the reasons for it to be one of the top-performing models.

Another method is *DANN*, an adversarial-based approach that combines a domain discriminator with a label classifier. The success of this approach is based on the ability to learn domain-invariant features between different domains in our skin lesion datasets. It is also the model that showed the highest improvement in performance for the challenging *MA* dataset. At this point, it is essential to emphasize that in our analysis, the top performing model *BSP* is a regularizer on top of *DANN*.

JAN is the leading moment matching approach performance-wise. Fundamentally it is learning transferable representations between our domains. We assume that this model worked better in comparison to other moment matching methods because it reduces the shift in joint distributions of the activation in the networks task-specific layers in comparison to matching the marginal distributions of features across domains. However *JAN* works on 9/10 datasets, unlike the above three methods that improved the performance on all available domains. *JAN* (0.14 AUPRC) did not seem to improve the performance for the challenging *MA* dataset in comparison to the unadapted baseline (0.15 AUPRC).

We were able to observe that the extension to the adversarial generative approach (*CDAN*) did not perform as well as purely adversarial methods. This might be due to the fact that *CDAN* is developed for aligning different domains of multi-modal distributions in an adversarial framework. Similarly to the performance of *JAN*, even *CDAN* failed to show an improvement in performance for *MA*. A different way to handle domain adaptation represents *AFN*, which showed performance improvement for *MA* dataset, but performance degradation for *HA*, where it is the least performing method in comparison to all other methods as shown in Fig. 2. This shows that progressively adapting the feature norms of specifically *HA* did not result in a transfer gain between the domains. Also the idea of utilizing less class confusion to imply more transferability in *MCC* did not seem to lead to an improvement in most of the domains. As seen in Fig. 2 and Table A.3, *MCC* is consistently one of the bad-performing methods in most of the domains.

Of all the models we experimented with, *DAN* is the least performing model for all domains. Also, as shown in Table A.3, *DAN* is the only model that did not work as the best performer to even at least one dataset. This shows that the architecture of DAN does not seem to be tailored for our skin lesion task.

Regarding the AUROC and AUPRC metrics, it appears that adversarial methods and their extensions generally outperform moment matching methods, with the exception of *JAN* for the *MLH* dataset in terms of AUPRC score. This finding is consistent with existing literature indicating that adversarial methods tend to perform better than moment matching methods [11].

4.3. Performance of UDA methods on non-dermoscopic datasets

One of the main reasons for the usage of AUROC for evaluation is that various works on UDA methods compare their results either with

Table A.1
Comparison of AUROC results across different derived datasets and UDA methods. The columns represent the domain shifted target datasets for the source dataset *H* (not listed here). Each row represents the results for a particular UDA method, with the first row indicating the results for the unadapted baseline method (*Src*). The best-performing UDA method for each dataset is highlighted in bold.

(UDA) method	Domain shifted dataset									
	HA	HLH	HLP	B	BA	BLH	BLP	M	MA	MLH
Src	0.65±0.04	0.74±0.05	0.82±0.08	0.65±0.01	0.58±0.05	0.61±0.03	0.62±0.07	0.51±0.01	**0.60±0.05**	0.57±0.03
DAN	0.64±0.06	0.79±0.02	0.85±0.06	0.67±0.02	0.59±0.02	0.65±0.02	0.70±0.06	0.52±0.02	0.52±0.05	0.59±0.04
JAN	0.71±0.04	0.85±0.03	0.92±0.02	0.76±0.01	0.69±0.01	0.74±0.02	0.69±0.04	0.62±0.03	0.55±0.06	0.61±0.02
DANN	0.73±0.03	0.84±0.03	0.91±0.03	**0.78±0.01**	0.67±0.02	0.75±0.01	0.72±0.02	0.62±0.01	**0.60±0.03**	0.62±0.03
ADDA	**0.74±0.06**	0.84±0.01	0.92±0.01	**0.78±0.01**	0.68±0.04	**0.77±0.01**	0.70±0.03	0.62±0.01	0.59±0.03	**0.63±0.03**
CDAN	0.71±0.05	0.85±0.01	0.90±0.03	0.77±0.01	0.68±0.03	0.75±0.02	0.71±0.02	**0.64±0.02**	0.56±0.02	0.62±0.03
BSP	0.72±0.03	0.84±0.01	**0.94±0.02**	**0.78±0.01**	**0.70±0.04**	0.75±0.02	0.71±0.02	**0.64±0.02**	0.56±0.02	0.62±0.02
AFN	0.67±0.03	0.85±0.02	0.92±0.05	0.76±0.01	0.66±0.01	0.73±0.01	**0.73±0.03**	0.60±0.01	0.55±0.02	0.60±0.02
MCC	0.70±0.01	**0.87±0.03**	**0.94±0.02**	0.76±0.01	0.68±0.03	0.73±0.01	0.70±0.04	0.61±0.03	0.58±0.08	0.59±0.02

AUROC or accuracy. In particular, in the medical field it is common practice to compare methods based on AUROC scores [70–72]. Most of the domain adaptation studies use accuracy as their metric for comparisons of methods, but none of these studies discuss the possible imbalance in their datasets.

Typical datasets used for domain adaptation tasks are, for instance, MNIST or Office-31. These images are easier to adapt to and differ a lot more than dermoscopic images do. Moreover, benchmark datasets for UDA are typically large, have almost balanced classes and the classification ability can even be validated by non-expert humans. On the contrary, dermoscopic images look very similar, making the task not only difficult for medical experts but also for the neural network. Backgrounds can contain unwanted complex structures used by the neural networks for training, such as black borders or hair.

When comparing all used methods, DAN performs poorly in our dermoscopic scenario, as well as in other adaptation tasks [10,11,18–20,24,39,41]. Our selection of UDA methods is benchmarked on Office-31-, Office-Home, ImageCLEF-DA-, and VisDA17-datasets. It is worth noting that not all UDA methods are compared in each work and on each of these datasets. Therefore, a fair comparison of non-dermoscopic results is not possible.

We can observe, that the adversarial UDA methods, namely *BSP* and *ADDA*, which are overall performing better in our dermoscopic scenario, also perform very well in other image classification tasks. According to the authors of [39], their method *BSP* specifically boosts the performance on relatively difficult tasks, where the source domain is quite small. *ADDA* was not often compared in these works, but it is outperformed by *CDAN* in the Office-31 adaptation. For all other methods, it is not possible to provide a clear order of performance improvements as they are compared on different tasks and with different methods. *DAN*, *DANN* and *JAN* are outperformed by all mentioned methods in Office-31-, ImageCLEF-DA-, Office-Home- and VisDA17-adaptation. *CDAN* is outperformed in Office-Home-, Office-31- and VisDA17-adaptation by *AFN*, *MCC* and *BSP+CDAN*.

5. Conclusion

This is the first work benchmarking UDA methods on dermoscopic datasets. We enable the reproducibility of results and their interpretations due to the utilization of publicly available datasets. Furthermore, the domain shifts between the derived datasets were quantified, unlike for most benchmark studies. Our analysis reveals that all selected UDA methods from different technical approaches improve the performance for most datasets compared to the unadapted baseline, however to different extents.

We have additionally performed a comparative analysis to examine how the performance of UDA methods is influenced by dataset- and method-specific factors, such as class imbalance and type of approach. It became evident that the overall performance of UDA methods depends on combinations of these factors.

Moreover, we compared the resulting performance of our selected UDA methods on dermoscopic images to the performance of other common benchmark adaptation tasks. In most cases, the UDA methods that performed well on dermoscopic datasets also proved to be the top performers in other non-dermoscopic tasks.

In our analysis, our aim was to compare different UDA methods to the same (well-established) ResNet-50 model as a baseline. However, the selection of the pre-trained model typically impacts the overall performance, which could be investigated in the future. Additionally, it is worth noting that noise in dermoscopic images may also affect a classifier's performance. Implementing denoising techniques for skin lesion classification could therefore be an interesting direction in the following studies. Investigating the intensity of performance degradation when gradually reducing the target dataset size or melanoma ratio should be investigated, too. Another possible area of interest is multi-source domain adaptation, where multiple modalities, e.g. clinical and dermoscopic images, are included in a skin lesion classifier. Our general recommendations favor the use of adversarial methods for UDA, as these consistently demonstrated substantial improvements. Ideally, the datasets should be large and balanced, because a low melanoma ratio was indicative of a high performance variance, thus making the performance of the applied methods uncertain.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Titus Josef Brinker would like to disclose that he is the owner of Smart Health Heidelberg GmbH (Handschuhsheimer Landstr. 9/1, 69120 Heidelberg, Germany) which develops mobile apps, outside of the submitted work.

Acknowledgments

This research is funded by the Helmholtz Artificial Intelligence Cooperation Unit[grant number ZT-I-PF-5-066].

The Helmholtz AI funding enabled the close cooperation between DKFZ and DLR, which leads to an interdisciplinary exchange between two research groups and thereby enables the integration of novel perspectives and experiences.

Appendix

See Tables A.1–A.3.

Table A.2
Comparison of change in AUROC results for each UDA method with respect to the unadapted baseline method (*Src*) for different datasets. H is the source dataset and the target datasets are listed in columns. The rows represent the improvements (in %) for each UDA-method.

		Domain shifted dataset									
		HA	HLH	HLP	B	BA	BLH	BLP	M	MA	MLH
UDA method	DAN	−1.54	6.76	3.66	3.08	1.72	6.56	12.9	1.96	−13.33	3.51
	JAN	9.23	14.86	12.20	16.92	18.97	21.31	11.29	21.57	−8.33	7.02
	DANN	12.31	13.51	10.98	**20.00**	15.52	22.95	16.13	21.57	**0**	8.77
	ADDA	**13.85**	13.51	12.20	**20.00**	17.24	**26.23**	12.90	21.57	−1.67	**10.53**
	CDAN	9.23	14.86	9.76	18.46	17.24	22.95	14.52	**25.49**	−6.67	8.77
	BSP	10.77	13.51	**14.63**	**20.00**	**20.69**	22.95	14.52	**25.49**	−6.67	8.77
	AFN	3.08	14.86	12.20	16.92	13.79	19.67	**17.74**	17.65	−8.33	5.26
	MCC	7.69	**17.57**	**14.63**	16.92	17.24	19.67	12.9	19.61	−3.33	3.51

Table A.3
Comparison of change in AUPRC results for each UDA method with respect to the unadapted baseline method (*Src*) for different datasets. H is the source dataset and the target datasets are listed in columns. The rows represent the improvements (in %) for each UDA-method.

		Domain shifted dataset									
		HA	HLH	HLP	B	BA	BLH	BLP	M	MA	MLH
UDA method	DAN	−14.29	11.59	27.03	5.26	5.26	6.85	7.79	8.82	−13.33	1.47
	JAN	7.14	18.84	51.35	26.32	78.95	16.44	6.49	29.41	−6.67	**7.35**
	DANN	21.43	17.39	48.65	29.82	68.42	16.44	9.09	29.41	**20.00**	5.88
	ADDA	**28.57**	17.39	48.65	29.82	**89.47**	**19.18**	7.79	29.41	13.33	**7.35**
	CDAN	0	18.84	45.95	28.07	73.68	16.44	9.09	**38.24**	−6.67	**7.35**
	BSP	14.29	18.84	**75.68**	**31.58**	78.95	17.81	7.79	35.29	13.33	**7.35**
	AFN	−21.43	**20.29**	54.05	28.07	84.21	15.07	**11.69**	26.47	6.67	4.41
	MCC	7.14	**20.29**	54.05	21.05	57.89	13.70	5.19	20.59	−6.67	4.41

References

[1] Barros Mendes Danilo, Correia da Silva Nilton. Skin Lesions Classification Using Convolutional Neural Networks in Clinical Images. 2018, http://dx.doi.org/10.48550/arXiv.1812.02316.

[2] Pious Ignatious K, Srinivasan R. A review on early diagnosis of skin cancer detection using deep learning techniques. In: 2022 international conference on computer, power and communications. IEEE; 2022, http://dx.doi.org/10.1109/iccpc55978.2022.10072274.

[3] Goceri Evgin. Automated skin cancer detection: Where we are and the way to the future. In: 2021 44th international conference on telecommunications and signal processing. 2021, p. 48–51. http://dx.doi.org/10.1109/TSP52935.2021.9522605.

[4] Yap Jordan, Yolland William, Tschandl Philipp. Multimodal skin lesion classification using deep learning. Exp Dermatol 2018;27(11):1261–7. http://dx.doi.org/10.1111/exd.13777.

[5] Bissoto Alceu, Fornaciali Michel, Valle Eduardo, Avila Sandra. (De)constructing bias on skin lesion datasets. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) workshops. 2019, http://dx.doi.org/10.48550/arXiv.1904.08818.

[6] Sun Qilin, Huang Chao, Chen Minjie, Xu Hui, Yang Yali. Skin lesion classification using additional patient information. BioMed Res Int 2021;2021:1–6. http://dx.doi.org/10.1155/2021/6673852.

[7] Goceri Evgin. Convolutional neural network based desktop applications to classify dermatological diseases. In: 2020 IEEE 4th international conference on image processing, applications and systems. 2020, p. 138–43. http://dx.doi.org/10.1109/IPAS50080.2020.9334956.

[8] Yosinski Jason, Clune Jeff, Bengio Yoshua, Lipson Hod. How transferable are features in deep neural networks? In: Advances in neural information processing systems. Vol. 27. 2014, http://dx.doi.org/10.48550/arXiv.1411.1792.

[9] Ovadia Yaniv, Fertig Emily, Ren Jie, Nado Zachary, Sculley D, Nowozin Sebastian, et al. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R, editors. Advances in neural information processing systems. Vol. 32. Curran Associates, Inc.; 2019, http://dx.doi.org/10.48550/arXiv.1906.02530.

[10] Tzeng Eric, Hoffman Judy, Saenko Kate, Darrell Trevor. Adversarial discriminative domain adaptation. In: 2017 IEEE conference on computer vision and pattern recognition. IEEE; 2017, http://dx.doi.org/10.1109/cvpr.2017.316.

[11] Xu Ruijia, Li Guanbin, Yang Jihan, Lin Liang. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In: 2019 IEEE/CVF international conference on computer vision. IEEE; 2019, http://dx.doi.org/10.1109/iccv.2019.00151.

[12] Yao Huaxiu, Wang Yu, Li Sai, Zhang Linjun, Liang Weixin, Zou James, et al. Improving out-of-distribution robustness via selective augmentation. In: International conference on machine learning. PMLR; 2022, p. 25407–37. http://dx.doi.org/10.48550/arXiv.2201.00299.

[13] Goceri Evgin. Medical image data augmentation: Techniques, comparisons and interpretations. Artif Intell Rev 2023;56(11):12561–605. http://dx.doi.org/10.1007/s10462-023-10453-z.

[14] Wang Jindong, Lan Cuiling, Liu Chang, Ouyang Yidong, Qin Tao. Generalizing to unseen domains: A survey on domain generalization. In: Proceedings of the thirtieth international joint conference on artificial intelligence. International Joint Conferences on Artificial Intelligence Organization; 2021, http://dx.doi.org/10.24963/ijcai.2021/628.

[15] Wang Mei, Deng Weihong. Deep visual domain adaptation: A survey. Neurocomputing 2018;312:135–53. http://dx.doi.org/10.1016/j.neucom.2018.05.083.

[16] Guo Lin Lawrence, Pfohl Stephen R, Fries Jason, Johnson Alistair, Posada Jose, Aftandilian Catherine, et al. Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. Cold Spring Harbor Laboratory; 2021, http://dx.doi.org/10.1101/2021.06.17.21259092.

[17] Quinonero-Candela Joaquin, Sugiyama Masashi, Schwaighofer Anton, Lawrence Neil D. Dataset shift in machine learning. Mit Press; 2008, http://dx.doi.org/10.7551/mitpress/9780262170055.001.0001.

[18] Long Mingsheng, Cao Yue, Wang Jianmin, Jordan Michael. Learning transferable features with deep adaptation networks. In: International conference on machine learning. PMLR; 2015, p. 97–105, URL https://dl.acm.org/doi/10.5555/3045118.3045130, retrieved on 12/08/23.

[19] Ganin Yaroslav, Lempitsky Victor. Unsupervised domain adaptation by backpropagation. In: International conference on machine learning. PMLR; 2015, p. 1180–9, URL https://dl.acm.org/doi/10.5555/3045118.3045244, retrieved on 12/08/23.

[20] Long Mingsheng, Cao Zhangjie, Wang Jianmin, Jordan Michael I. Conditional adversarial domain adaptation. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. Advances in neural information processing systems. Vol. 31. Curran Associates, Inc.; 2018, https://proceedings.neurips.cc/paper/2018/file/ab88b15733f543179858600245108dd8-Paper.pdf, retrieved on 12/08/23.

[21] Pan Sinno Jialin, Yang Qiang. A survey on transfer learning. IEEE Trans Knowl Data Eng 2010;22(10):1345–59. http://dx.doi.org/10.1109/TKDE.2009.191.

[22] Patel Vishal M, Gopalan Raghuraman, Li Ruonan, Chellappa Rama. Visual domain adaptation: A survey of recent advances. IEEE Signal Process Mag 2015;32(3):53–69. http://dx.doi.org/10.1109/msp.2014.2347059.

[23] Zhang Youshan. A survey of unsupervised domain adaptation for visual recognition. 2021, http://dx.doi.org/10.48550/arXiv.2112.06745.

[24] Jin Ying, Wang Ximei, Long Mingsheng, Wang Jianmin. Minimum class confusion for versatile domain adaptation. In: Computer vision – ECCV 2020. Springer International Publishing; 2020, p. 464–80. http://dx.doi.org/10.1007/978-3-030-58589-1_28.

[25] Guan Hao, Liu Mingxia. Domain adaptation for medical image analysis: A survey. IEEE Trans Biomed Eng 2022;69(3):1173–85. http://dx.doi.org/10.1109/tbme.2021.3117407.

[26] Ayan Enes, Ünver Halil Murat. Data augmentation importance for classification of skin lesions via deep learning. In: 2018 electric electronics, computer science, biomedical engineerings' meeting. 2018, p. 1–4. http://dx.doi.org/10.1109/EBBT.2018.8391469.

[27] Pham Tri-Cong, Luong Chi-Mai, Visani Muriel, Hoang Van-Dung. Deep CNN and data augmentation for skin lesion classification. In: Intelligent information and database systems. Springer International Publishing; 2018, p. 573–82. http://dx.doi.org/10.1007/978-3-319-75420-8_54.

[28] Goceri Evgin. Comparison of the impacts of dermoscopy image augmentation methods on skin cancer classification and a new augmentation method with wavelet packets. Int J Imaging Syst Technol 2023;33(5):1727–44. http://dx.doi.org/10.1002/ima.22890.

[29] Yoon Chris, Hamarneh Ghassan, Garbi Rafeef. Generalizable feature learning in the presence of data bias and domain class imbalance with application to skin lesion classification. Lecture notes in computer science, Springer International Publishing; 2019, p. 365–73. http://dx.doi.org/10.1007/978-3-030-32251-9_40.

[30] Bissoto Alceu, Barata Catarina, Valle Eduardo, Avila Sandra. Artifact-based domain generalization of skin lesion models. Lecture Notes in Computer Science, Springer Nature Switzerland; 2023, p. 133–49. http://dx.doi.org/10.1007/978-3-031-25069-9_10.

[31] Fogelberg Katharina, Chamarthi Sireesha, Maron Roman C, Niebling Julia, Brinker Titus J. Domain shifts in dermoscopic skin cancer datasets: Evaluation of essential limitations for clinical translation. New Biotechnol 2023;76:106–17. http://dx.doi.org/10.1016/j.nbt.2023.04.006.

[32] Ben-David Shai, Blitzer John, Crammer Koby, Kulesza Alex, Pereira Fernando, Vaughan Jennifer Wortman. A theory of learning from different domains. Mach Learn 2009;79(1–2):151–75. http://dx.doi.org/10.1007/s10994-009-5152-4.

[33] Huang Jiayuan, Gretton Arthur, Borgwardt Karsten, Schölkopf Bernhard, Smola Alex. Correcting sample selection bias by unlabeled data. In: Schölkopf B, Platt J, Hoffman T, editors. Advances in neural information processing systems. Vol. 19. MIT Press; 2006, URL https://proceedings.neurips.cc/paper/2006/file/a2186aa7c086b46ad4e8bf81e2a3a19b-Paper.pdf, retrieved on 12/08/23.

[34] Sugiyama Masashi, Nakajima Shinichi, Kashima Hisashi, Buenau Paul, Kawanabe Motoaki. Direct importance estimation with model selection and its application to covariate shift adaptation. In: Platt J, Koller D, Singer Y, Roweis S, editors. Advances in neural information processing systems. Vol. 20. Curran Associates, Inc.; 2007, URL https://proceedings.neurips.cc/paper/2007/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper.pdf, retrieved on 12/08/23.

[35] Pan Sinno Jialin, Tsang Ivor W, Kwok James T, Yang Qiang. Domain adaptation via transfer component analysis. IEEE Trans Neural Netw 2010;22(2):199–210. http://dx.doi.org/10.1109/TNN.2010.2091281.

[36] Gong Boqing, Grauman Kristen, Sha Fei. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In: International conference on machine learning. PMLR; 2013, p. 222–30, URL https://proceedings.mlr.press/v28/gong13.html, retrieved on 12/08/23.

[37] Zhang Kun, Schölkopf Bernhard, Muandet Krikamol, Wang Zhikun. Domain adaptation under target and conditional shift. In: Proceedings of the 30th international conference on machine learning. Proceedings of machine learning research, vol.28, (3):PMLR; 2013, p. 819–27, URL https://proceedings.mlr.press/v28/zhang13d.html, retrieved on 12/08/23.

[38] Courty Nicolas, Flamary Rémi, Habrard Amaury, Rakotomamonjy Alain. Joint distribution optimal transportation for domain adaptation. In: Advances in neural information processing systems. Vol. 30. 2017, http://dx.doi.org/10.48550/arXiv.1705.08848.

[39] Chen Xinyang, Wang Sinan, Long Mingsheng, Wang Jianmin. Transferability vs. Discriminability: Batch spectral penalization for adversarial domain adaptation. In: Chaudhuri Kamalika, Salakhutdinov Ruslan, editors. Proceedings of the 36th international conference on machine learning. Proceedings of machine learning research, vol.97, PMLR; 2019, p. 1081–90, URL https://proceedings.mlr.press/v97/chen19i.html, retrieved on 12/08/23.

[40] Zellinger Werner, Grubinger Thomas, Lughofer Edwin, Natschläger Thomas, Saminger-Platz Susanne. Central moment discrepancy (cmd) for domain-invariant representation learning. 2017, http://dx.doi.org/10.48550/arXiv.1702.08811.

[41] Long Mingsheng, Zhu Han, Wang Jianmin, Jordan Michael I. Deep transfer learning with joint adaptation networks. In: International conference on machine learning. PMLR; 2017, p. 2208–17, URL https://dl.acm.org/doi/10.5555/3305890.3305909, retrieved on 12/08/23.

[42] Sun Baochen, Saenko Kate. Deep coral: Correlation alignment for deep domain adaptation. In: Computer vision–ECCV 2016 workshops. Springer; 2016, p. 443–50, URL http://dx.doi.org/10.1007/978-3-319-49409-8_35.

[43] Zhao Sicheng, Yue Xiangyu, Zhang Shanghang, Li Bo, Zhao Han, Wu Bichen, et al. A review of single-source deep unsupervised visual domain adaptation. IEEE Trans Neural Netw Learn Syst 2022;33(2):473–93. http://dx.doi.org/10.1109/tnnls.2020.3028503.

[44] Saito Kuniaki, Watanabe Kohei, Ushiku Yoshitaka, Harada Tatsuya. Maximum classifier discrepancy for unsupervised domain adaptation. In: 2018 IEEE/CVF conference on computer vision and pattern recognition. IEEE; 2018, http://dx.doi.org/10.1109/cvpr.2018.00392.

[45] Goodfellow Ian, Pouget-Abadie Jean, Mirza Mehdi, Xu Bing, Warde-Farley David, Ozair Sherjil, et al. Generative adversarial nets. In: Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger KQ, editors. Advances in neural information processing systems. Vol. 27. Curran Associates, Inc.; 2014, URL https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf, retrieved on 12/08/23.

[46] Arjovsky Martin, Bottou Leon. Towards principled methods for training generative adversarial networks. In: International conference on learning representations. 2017, URL https://openreview.net/forum?id=Hk4_qw5xe, retrieved on 12/08/23.

[47] Arora Sanjeev, Ge Rong, Liang Yingyu, Ma Tengyu, Zhang Yi. Generalization and equilibrium in generative adversarial nets (GANs). In: Precup Doina, Teh Yee Whye, editors. Proceedings of the 34th international conference on machine learning. Proceedings of machine learning research, vol.70, PMLR; 2017, p. 224–32, URL https://proceedings.mlr.press/v70/arora17a.html, retrieved on 12/08/23.

[48] Mirza Mehdi, Osindero Simon. Conditional generative adversarial nets. 2014, http://dx.doi.org/10.48550/arXiv.1411.1784.

[49] Liu Ming-Yu, Breuel Thomas, Kautz Jan. Unsupervised image-to-image translation networks. 2017, http://dx.doi.org/10.48550/arXiv.1703.00848.

[50] Sankaranarayanan S, Balaji Y, Castillo CD, Chellappa R. Generate to adapt: Aligning domains using generative adversarial networks. In: 2018 IEEE/CVF conference on computer vision and pattern recognition. Los Alamitos, CA, USA: IEEE Computer Society; 2018, p. 8503–12. http://dx.doi.org/10.1109/CVPR.2018.00887.

[51] Hoffman Judy, Tzeng Eric, Park Taesung, Zhu Jun-Yan, Isola Phillip, Saenko Kate, et al. Cycada: Cycle-consistent adversarial domain adaptation. In: International conference on machine learning. Pmlr; 2018, p. 1989–98, URL http://proceedings.mlr.press/v80/hoffman18a/hoffman18a.pdf, retrieved on 12/08/23.

[52] Gu Yanyang, Ge Zongyuan, Bonnington C Paul, Zhou Jun. Progressive transfer learning and adversarial domain adaptation for cross-domain skin disease classification. IEEE J Biomed Health Inform 2020;24(5):1379–93. http://dx.doi.org/10.1109/jbhi.2019.2942429.

[53] Ahn Euijoon, Kumar Ashnil, Fulham Michael, Feng Dagan, Kim Jinman. Unsupervised domain adaptation to classify medical images using zero-bias convolutional auto-encoders and context-based feature augmentation. IEEE Trans Med Imaging 2020;39(7):2385–94. http://dx.doi.org/10.1109/tmi.2020.2971258.

[54] Ringwald Tobias, Stiefelhagen Rainer. Adaptiope: A modern benchmark for unsupervised domain adaptation. In: 2021 IEEE winter conference on applications of computer vision. WACV, IEEE; 2021, http://dx.doi.org/10.1109/wacv48630.2021.00015.

[55] Peng Xingchao, Usman Ben, Kaushik Neela, Wang Dequan, Hoffman Judy, Saenko Kate. VisDA: A synthetic-to-real benchmark for visual domain adaptation. In: 2018 IEEE/CVF conference on computer vision and pattern recognition workshops. IEEE; 2018, http://dx.doi.org/10.1109/cvprw.2018.00271.

[56] Nagananda Navya, Taufique Abu Md Niamul, Madappa Raaga, Jahan Chowdhury Sadman, Minnehan Breton, Rovito Todd, et al. Benchmarking domain adaptation methods on aerial datasets. Sensors 2021;21(23):8070. http://dx.doi.org/10.3390/s21238070.

[57] Saat Parisa, Nogovitsyn Nikita, Hassan Muhammad Yusuf, Ganaie Muhammad Athar, Souza Roberto, Hemmati Hadi. A domain adaptation benchmark for T1-weighted brain magnetic resonance image segmentation. Front Neuroinform 2022;16. http://dx.doi.org/10.3389/fninf.2022.919779.

[58] Cassidy Bill, Kendrick Connah, Brodzicki Andrzej, Jaworek-Korjakowska Joanna, Yap Moi Hoon. Analysis of the ISIC image datasets: Usage, benchmarks and recommendations. Med Image Anal 2022;75:102305. http://dx.doi.org/10.1016/j.media.2021.102305.

[59] Tschandl Philipp, Rosendahl Cliff, Kittler Harald. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data 2018;5(1):1–9. http://dx.doi.org/10.1038/sdata.2018.161.

[60] Combalia Marc, Codella Noel CF, Rotemberg Veronica, Helba Brian, Vilaplana Veronica, Reiter Ofer, et al. BCN20000: Dermoscopic lesions in the wild. 2019, http://dx.doi.org/10.48550/arXiv.1908.02288.

[61] Jiang Junguang, Shu Yang, Wang Jianmin, Long Mingsheng. Transferability in deep learning: A survey. 2022, http://dx.doi.org/10.48550/arXiv.2201.05867, arXiv:2201.05867.

[62] He Kaiming, Zhang Xiangyu, Ren Shaoqing, Sun Jian. Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition. IEEE; 2016, http://dx.doi.org/10.1109/cvpr.2016.90.

[63] Jiang Junguang, Chen Baixu, Fu Bo, Long Mingsheng. Transfer-learning-library. 2020, GitHub repository, GitHub, https://github.com/thuml/Transfer-Learning-Library, retrieved on 12/08/23.

[64] Saito Takaya, Rehmsmeier Marc. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015;10(3):e0118432, URL http://dx.doi.org/10.1371/journal.pone.0118432.

[65] Zhang Tianran, Chen Muhao, Bui Alex AT. AdaDiag: Adversarial domain adaptation of diagnostic prediction with clinical event sequences. J Biomed Inform 2022;134:104168. http://dx.doi.org/10.1016/j.jbi.2022.104168.

[66] Zhang Luxin, Germain Pascal, Kessaci Yacine, Biernacki Christophe. Interpretable domain adaptation for hidden subdomain alignment in the context of pre-trained source models. In: Proceedings of the AAAI conference on artificial intelligence. Vol. 36, no. 8. Association for the Advancement of Artificial Intelligence (AAAI); 2022, p. 9057–65. http://dx.doi.org/10.1609/aaai.v36i8.20890.

[67] Duarte Ana F, Sousa-Pinto Bernardo, Azevedo Luís F, Barros Ana M, Puig Susana, Malvehy Josep, et al. Clinical ABCDE rule for early melanoma detection. Eur J Dermatol 2021;31(6):771–8. http://dx.doi.org/10.1684/ejd.2021.4171.

[68] Scope Alon, Marchetti Michael A, Marghoob Ashfaq A, Dusza Stephen W, Geller Alan C, Satagopan Jaya M, et al. The study of nevi in children: Principles learned and implications for melanoma diagnosis. J Am Acad Dermatol 2016;75(4):813–23. http://dx.doi.org/10.1016/j.jaad.2016.03.027.

[69] Motiian Saeid, Jones Quinn, Iranmanesh Seyed Mehdi, Doretto Gianfranco. Few-shot adversarial domain adaptation. In: Proceedings of the 31st international conference on neural information processing systems. Red Hook, NY, USA: Curran Associates Inc.; 2017, p. 6673–83, URL https://dl.acm.org/doi/10.5555/3295222.3295412, retrieved on 12/08/23.

[70] Purushotham Sanjay, Carvalho Wilka, Nilanon Tanachat, Liu Yan. Variational recurrent adversarial deep domain adaptation. In: International conference on learning representations. 2017, URL https://openreview.net/forum?id=rk9eAFcxg, retrieved on 12/08/23.

[71] Zhou Jieli, Jing Baoyu, Wang Zeya, Xin Hongyi, Tong Hanghang. SODA: Detecting COVID-19 in chest X-Rays with semi-supervised open set domain adaptation. IEEE/ACM Trans Comput Biol Bioinform 2022;19(5):2605–12. http://dx.doi.org/10.1109/tcbb.2021.3066331.

[72] Feng Yangqin, Wang Zizhou, Xu Xinxing, Wang Yan, Fu Huazhu, Li Shaohua, et al. Contrastive domain adaptation with consistency match for automated pneumonia diagnosis. Med Image Anal 2023;83:102664. http://dx.doi.org/10.1016/j.media.2022.102664.