Domain Adaptation for Skin Lesion: Evaluating Real-World Generalisation

Nurjahan Sultana, Wenqi Lu, Xinqi Fan, Moi Hoon Yap
Department of Computing and Mathematics, Manchester Metropolitan University
Dalton Building, Chester Street, M1 5GD Manchester, UK
nurjahan.sultana@stu.mmu.ac.uk; {w.lu, x.fan, m.yap}@mmu.ac.uk

Abstract

Domain shifts limit the generalisation of deep learning models for skin cancer detection, particularly when trained on dermoscopic images but deployed on clinical images. This study evaluates supervised and unsupervised domain adaptation techniques to improve model performance on a diverse set of clinical images. We introduce the IMPS dataset, a varied collection of clinical skin lesion images, to assess robustness under real-world conditions. Experimental results show that unsupervised methods, particularly Domain-Adversarial Neural Networks (DANN), provide better generalisation than supervised approaches. These findings suggest that evaluating models on limited datasets may give an incomplete picture of their reliability. Future research should test these approaches on additional clinical datasets that were not part of this study to better assess their suitability for real-world applications. Our GitHub repository contains the IMPS dataset and image IDs referencing the original dataset sources: https://github.com/mmu-dermatology-research/sl_domain_adaptation

1. Introduction

Skin cancer is the third most common human malignancy and rising globally at an alarming rate [19]. For many years, researchers have been working to develop deep learning models that can effectively detect skin cancer in its early stages. While many Convolutional Neural Network (CNN) models have demonstrated superior performance in skin cancer classification compared to dermatologists in experimental settings, their generalisation issues have prevented their implementation in clinics [14]. The performance difference mainly arises due to machine learning models assumption that the training dataset and test dataset share the same data distribution [21]. However, in real-world scenarios, models often encounter different types of images than those they were trained on or evaluated with. The prediction on unseen data can be less accurate or unreliable due to even a small-scale deviation from the distribution of the training domain [54], [38], [48].

This differing distribution of the training dataset and the real-world dataset has been a major obstacle to skin cancer classification. The three primary imaging modalities used to diagnose skin diseases are clinical images, dermoscopic images, and histopathological images [52]. Clinical images are pictures of skin problems taken by doctors with their phones to examine them and keep them in patients' files [18]. Dermoscopic images, on the other hand, are taken by using a dermoscope, which is a non-invasive handheld skin imaging device that uses optical magnification and cross-polarised lighting [55]. They are ideal for computer-aided diagnosis of skin lesions due to their clear, well-lit close-up views with minimal background interference [45]. This makes them easier to process for computer vision analysis compared to traditional clinical images. However, in situations where access to a dermoscope is limited, doctors rely on images taken by their mobile phones for an initial assessment. These images have different angles, lighting, colour brightness, and many other variations compared to dermoscopic images. This deviation between datasets is commonly referred to as a domain shift [5]. Neural networks excel at fitting data but struggle with generalisation to unseen data, particularly when differences in image acquisition between medical centers, are present [43]. Maron et al. [35] establish a benchmark to assess the robustness of classifiers against out-of-distribution data and their findings indicate the vulnerabilities of CNN to these challenges.

Researchers have tried to address the issue of domain shifting using techniques such as domain adaptation. Domain adaptation tackles challenges where training data (source domain) differs from real-world application (target domain). It bridges this gap by aligning the data distributions so that the trained model performs well in the target domain [13]. However, to properly assess the effectiveness of domain adaptation, the selection criteria for source and target domains are crucial. Chamarthi et al. [4] examined technical and biological shifts between source and target datasets in applying domain adaptation methods in skin

cancer classification. Others have used different datasets as different domains to experiment domain adaptation [50], [14], [20]. However, there is a gap in the research regarding how domain adaptation techniques can enhance models trained on the widely available modality, dermoscopic images, and improve model generalisation on datasets from other modalities and multiple sources which reflect real-world variations.

Our primary contribution is a systematic comparison of state-of-the-art supervised and unsupervised domain adaptation methods for skin cancer classification under real-world conditions. In most studies, the target domain is usually not significantly more diverse than the source. In contrast, we examine not only a standard source–target scenario but also a target domain deliberately constructed to be more heterogeneous than the source, allowing us to assess generalisation under more realistic conditions. This design mirrors the practical scenario in which models trained on carefully controlled (dermoscopic) data must handle a wider range of clinical images. It also exposes how assessments conducted on less diverse target sets may overestimate model robustness, thereby emphasising the critical importance of evaluating under realistic conditions.

Our second contribution is the curation of a diverse dataset to use as the target domain. We aim to evaluate the effectiveness of domain adaptation in skin cancer classification models when the distribution of the target domain is more varied than that of the source domain. Our source domain consists of dermoscopic images from the International Skin Imaging Collaboration (ISIC) archive. To ensure our target domain mimics real-world scenarios, we created a new dataset named IMPS by combining clinical images from SD198 [45], ISIC Archive Gallery [28], MED-NODE [16], and PAD-UFES-20 [39]. These sources were carefully selected to ensure significant variation in demographics, lighting conditions, and image acquisition techniques. For instance, SD198 consists of images downloaded from DermQuest, recognised and labeled by experts, reflecting real-world variations in colour, exposure, and illumination. MED-NODE includes images taken with digital cameras, introducing variability in image quality and conditions, which mirrors practical clinical challenges. PAD-UFES-20 features images from the Dermatological and Surgical Assistance Program at the Federal University of Espírito Santo in Brazil, representing different demographic backgrounds. By ensuring a significant difference between source and target domains, we carried out cross-domain evaluation experiments where we trained deep learning models on dermoscopic images and evaluated their performance on the IMPS dataset. This serves as a benchmark for the performance of the models in more diverse scenarios.

2. Related Work

2.1. Supervised Domain Adaptation Methods

Early supervised domain adaptation (SDA) methods often require a small number of labelled target samples to guide the alignment of feature distributions across source and target domains. Tzeng et al. [46] addressed domain bias by proposing a CNN architecture that reduces marginal distribution discrepancies and preserves class relationships via a label distribution matching loss. Although they also explored semi-supervised settings, their main approach demonstrates how some target labels can improve transfer, using Office and Caltech-256 datasets. Their target sets do not necessarily exceed the diversity of their source sets, but the method scales to multi-domain benchmarks.

Goetz et al. [17] tackled sparse annotations in MRI brain tumour segmentation (the "target"), relying on reweighting to counter sampling bias. Although their target data were not necessarily more diverse than the source, it exemplifies how even limited labelled target samples can suffice under a supervised paradigm. Likewise, Liu et al. [34] combined image and feature adaptation in a supervised framework for cross-domain change detection in bi-temporal remote-sensing images, aligning labelled pre- and post-event images. Their target datasets (e.g. CDD) vary in conditions (seasonal and illumination changes), but both source and target contain labelled images.

Hedegaard et al. [25] approached SDA by framing it as a Graph Embedding problem (DAGE), using labelled source and target samples to learn a domain-invariant latent space with improved generalisation in Office31, Digits, and VisDA. Motiian et al. [36] proposed a Siamese-network-based Classification and Contrastive Semantic Alignment (CCSA) loss, also requiring a few labelled target samples, tested on Office31, MNIST-USPS, and VLCS. Although these works do not explicitly focus on a target domain that is more diverse than the source, they do show that even a small subset of labelled target data can significantly improve transfer performance.

In the medical domain, Carretero et al. [1] developed a Supervised Contrastive Domain Adaptation (SCDA) strategy for histopathological images. They used labelled samples from multiple hospitals, thus dealing with multi-centre data shifts. Although each hospital domain contains similar types of lesions, consolidating them can lead to a target domain that is arguably more heterogeneous than a single-hospital source. Their results showed a notable jump in balanced accuracy even with as few as 8–10 labelled slides per class.

While Huang et al. [27] focused on comparing semi-supervised and self-supervised learning in medical imaging, their semi-supervised methods (e.g. MixMatch) effectively treat labelled and unlabelled data within the target domain.

Strictly speaking, these approaches still benefit from partial supervision on the target side, fitting within the broader scope of supervised or semi-supervised domain adaptation.

2.2. Unsupervised Domain Adaptation Methods

Unsupervised domain adaptation (UDA) methods do not require target labels, making them attractive in scenarios where annotation is prohibitively expensive. Early seminal approaches include the Gradient Reversal Layer of Ganin and Lempitsky [15] and ADDA by Tzeng et al. [47], both aligning feature distributions adversarially. Sun and Saenko [44] introduced Deep CORAL to match second-order statistics, while Xie et al. [53] proposed MSTN to align semantic representations. These methods mostly focus on balanced source–target domains (Office, MNIST, etc.) rather than a target domain that is strictly more diverse.

In medical imaging, many UDA works tackle settings where target labels are unavailable or scarce. Chen et al. [6] used Generative Adversarial Networks to bridge CT and MR images in cross-modality segmentation (SIFA). Hou et al. [26] proposed DASQE for quality enhancement without reference images, while Omidi et al. [37] applied adversarial methods to adapt adult MRI skull-stripping models to neonatal data—here, the neonatal target domain is distinct but not necessarily "larger" or "more diverse." De Bel et al. [11] addressed stain variation in renal histopathology by unpaired image-to-image translation, showing how even small shifts in domain can degrade performance without adaptation.

Several studies on skin lesion classification employed UDA to handle domain shifts caused by different acquisition devices or patient populations [3, 14, 49]. Chamarthi et al. [3] benchmarked multiple UDA methods on eleven dermoscopic datasets, noting improvements when the target domain differs in both technical and biological factors. In some instances, the target domain (e.g. images from new hospitals or novel imaging devices) turned out to be more variable than the limited source set, mirroring the real-world scenario of encountering diverse clinical images after training on controlled data. Wang et al. [49] further explored multi-source UDA for fairer skin lesion classification across demographic groups, emphasising how combining various source domains can help accommodate an even broader target.

Methods such as CODA [24] or DCAN [32] similarly show large performance gains under shifting imaging conditions without labelled target data. CODA, for instance, adapts a feature extractor online to out-of-domain samples in High Content Imaging; while the source set is carefully controlled, the unlabelled target data come from multiple pharmaceutical labs with disparate equipment, thus more diverse. Similarly, DCAN [32] introduced domain-conditioned channel attention and feature correction blocks to tackle large-scale domain shifts in Office-Home and DomainNet. These examples highlight that, when the target domain has greater variability (e.g. multiple labs, more patient demographics), carefully designed adversarial and alignment strategies are especially beneficial.

Other notable UDA approaches include source-free adaptation [31] (where the original source data cannot be shared for privacy reasons) and multi-source methods [42], which exploit multiple pre-trained source models. Surveys by Kumari and Singh [30] and Guan and Liu [22] provide broader overviews, categorising UDA approaches into feature alignment, image translation, disentangled representations, self-supervised methods, and more. A recurring theme is that the more diverse the unlabelled target domain, the greater the benefit from robust feature-invariance strategies and regularisation to handle out-of-distribution samples.

3. Methodology

3.1. Dataset

In this study, we used ISIC Curated Balanced Dataset [2], [23], [8], [7], [9], [41] as the source domain. This dataset includes 9810 dermoscopic images, with 7848 images in the training set (3924 melanoma and 3924 other cases) and 1962 images in the validation set (981 melanoma and 981 other). However, for clinical images of skin lesions, there is currently no dataset with diverse scenes. Therefore, we built the IMPS dataset, comprising 1,657 images (828 melanoma and 830 other skin conditions), which is divided into training (1,159), validation (129), and test (249) sets. This collection integrates several datasets: SD198, ISIC (clinical), MED-NODE and PAD-UFES-20. The SD198 dataset focuses on malignant melanoma and lentigo maligant melanoma (combined into a single class "Melanoma") and various other skin cancers. Hence, we collected 373 melanoma and 375 other images from SD198. ISIC clinical images, from which 512 images were curated out of 530 images. MED-NODE, with 170 images across two classes (melanoma and other). PAD-UFES-20, from which

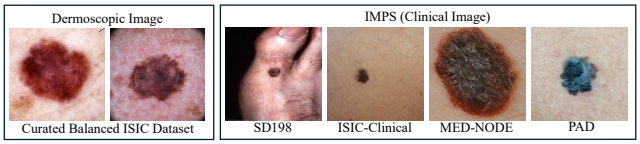

Figure 1. The image compares two domains of skin lesion imaging datasets: dermoscopic images from the "Curated Balanced ISIC Dataset" (left) showing magnified, specialized views of lesions, and clinical images (right) from four different datasets (SD198, ISIC-Clinical, MED-NODE, and PAD) representing standard clinical photography of various skin lesions with different appearances and characteristics.

52 (melanoma) images and 280 (other) images from five other skin cancer classes were selected using stratified sampling. The source and composition of IMPS dataset will be made available upon request.

Figure 1 shows the example images from each domain and Figure 2 illustrates distinct differences in pixel intensity distributions between dermoscopic images from the Curated balanced ISIC dataset and clinical images from the IMPS dataset. Dermoscopic images exhibit a wider range of pixel intensities, suggesting richer colour detail and enhanced diagnostic potential. In contrast, clinical images show narrower intensity distributions, indicative of color loss and diminished image detail.

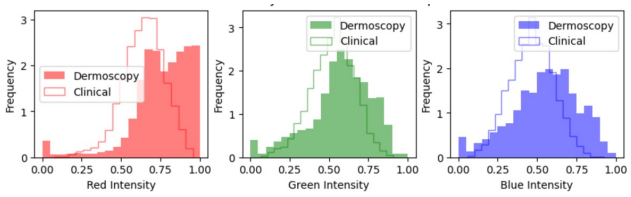

Figure 2. The histograms display the pixel intensity distributions across red, green, and blue channels for 100 images each from the dermoscopic and clinical domains. The dermoscopy images are represented by solid colours, and the clinical images, which undergo artificial degradation to simulate domain shift, are shown with outlined bars. Each histogram ranges from 0 to 1, representing normalized pixel values. This visualization facilitates a comparative analysis of colour profiling between the two domains.

Table 1. Description of the IMPS dataset and its source clinical images from the SD198 [45], ISIC [28], MED-NODE [16], and PAD-UFES-20 [39]. Diversity: variations in colour brightness, picture angles, demographics, skin tones, and image acquisition conditions, all of which underscore its real-world heterogeneity and applicability.

Dataset	Diversity	Total Number	Total Class	Tools Used
SD198	Comprehensive	6,584	198	Images collected from DermQuest
ISIC	High	530	4	Digital Cameras
MED-NODE	Limited	170	2	DSLR (Nikon D3 & D1x)
PAD-UFES-20	Moderate	2,298	7	Smartphone
IMPS	High	1657	2	Smartphone and Digital Cameras

Table 1 summarises the proposed IMPS dataset, which exhibits exceptional diversity as its images are captured across different regions using a variety of devices.This results in notable variations in colour brightness, image angles, and acquisition conditions, and the dataset also encompasses a wide range of demographic attributes reflecting diverse skin tones and age groups which collectively enhance its heterogeneity and real-world relevance. As illustrated in Figure 3, all four t-SNE plots show that each dataset occupies a distinct region in feature space, despite representing the same broader clinical domain. The density, spread, and relative clustering of Melanoma (red) vs. Other (blue) points vary considerably, suggesting that these datasets differ in ways that could affect classification consistency.

It is noteworthy to mention that the images are preprocessed by resizing them while maintaining their aspect ratio, specifically shortening the side to fit within a 224×224 resolution. All images are checked for duplicates and removed during the curation process. The IMPS classes Melanoma ("mel") and Other ("oth") are taken to resemble the Balanced dataset classes. The IMPS dataset is balanced based on the total number of "mel" and "oth" images. As previ-

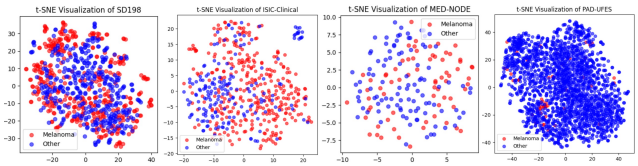

Figure 3. From left to right, the t-SNE plots for (SD198, ISIC-Clinical, MED-NODE, and PAD-UFES) clinical dataset, illustrate the distribution of extracted features. Red points correspond to Melanoma, and blue points denote Other skin lesions, highlighting the distinct clustering patterns across each dataset.

ously discussed, we selected the source and target datasets in a way that ensures a difference in their distributions. To confirm the statistical significance of the difference, we conducted a Kolmogorov–Smirnov test [12] under the null hypothesis that both the target and source datasets are drawn from the same distribution. We found the p-value to be 1.54×10^{-13}, which is much smaller than any commonly used significance level. Therefore, we conclude that there is a significant difference between the source and target domains.

3.2. Method

Base Model: We employ EfficientNet-B2 as the backbone for our base model, selected for its state-of-the-art performance and favourable computational efficiency in image classification tasks. EfficientNet models use a compound scaling approach to optimise depth, width, and resolution, enhancing feature extraction for medical imaging. This makes them well-suited for handling variations in texture, colour, and structure, which are critical for accurate diagnosis. Their balanced architecture improves precision while maintaining computational efficiency, outperforming traditional CNNs in both accuracy and resource usage [29]. The model is trained on a balanced ISIC dermoscopic dataset—leveraging high-quality labels—to learn discriminative feature representations, and is then evaluated on the IMPS dataset as well as on each subset (I, M, P, S) separately to establish baseline performance metrics.

Supervised Domain Adaptation (SDA): SDA adapts a model from a labelled source domain to a labelled target domain by leveraging labels from both domains for im-

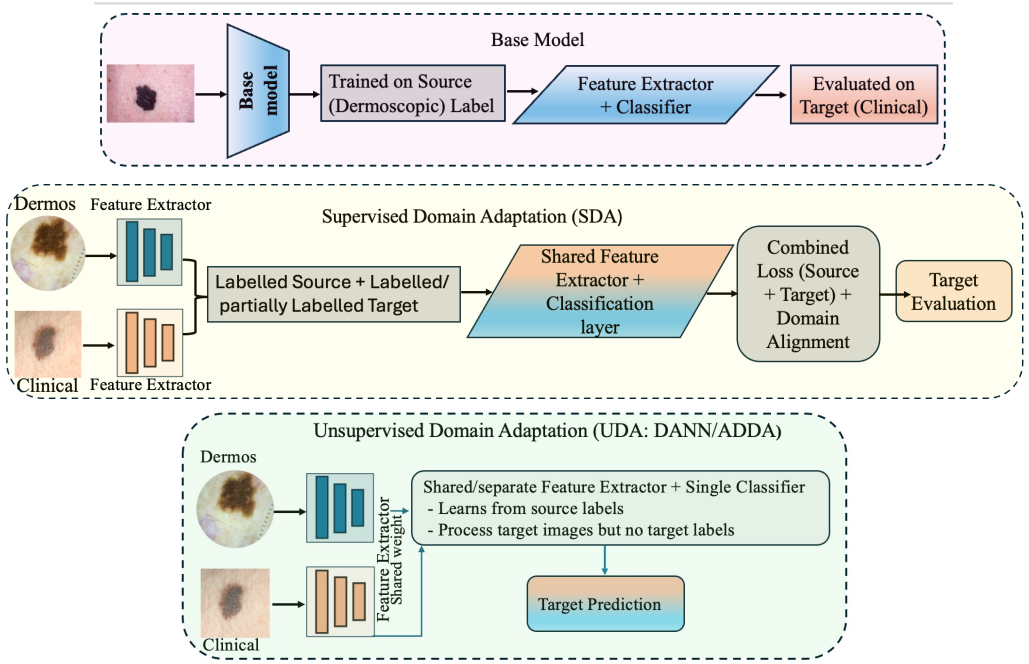

Figure 4. We begin by training EfficientNet-B2 on a balanced ISIC dermoscopic dataset as the source domain (Base Model), then evaluate on IMPS (clinical dataset) and each of its constituent subsets (I, M, P, S). For Supervised Domain Adaptation (SDA), we employ ATDOC and LIC, using the same balanced dermoscopic data as source alongside labelled and clinical data from IMPS to learn domain-invariant representations. For Unsupervised Domain Adaptation (UDA), we adopt DANN and ADDA, again with the balanced dermoscopic data as source but unlabelled clinical images as target. All methods share the same training and evaluation protocol, facilitating a direct comparison of SDA and UDA performance across IMPS and its individual subsets.

proved performance. Unlike Unsupervised Domain Adaptation (UDA), which relies only on unlabelled target data, SDA enables direct learning from target labels while retaining knowledge from the source [10].

The Auxiliary Target Domain-Oriented Classifier (ATDOC) is a Supervised Domain Adaptation (SDA) method designed to reduce classifier bias when adapting from a source to a target domain. It introduces an auxiliary classifier for target data, generating unbiased pseudo labels through non-parametric classifiers: the Nearest Centroid Classifier (NC) and Neighborhood Aggregation (NA). NC assigns labels using class centroids stored in a memory bank, while NA refines labels using nearest neighbours. The final confidence-weighted pseudo-labeling loss is defined as:

$$L_{\text{ATDOC-NA}} = -\frac{\lambda}{N_{tu}} \sum_{i=1}^{N_{tu}} \hat{q}_{i,\hat{y}_i} \log p_{i,\hat{y}_i} \tag{1}$$

where $\hat{q}_{i,\hat{y}_i}$ is the confidence-weighted pseudo-label and $p_{i,\hat{y}_i}$ is the predicted probability [33].

Supervised Adapters for Domain Adaptation in Learned Image Compression (LIC) introduce domain-specific adapters at the decoder, each dedicated to a target domain along with one for the source. A gate network predicts domain probabilities and blends adapter outputs to enhance reconstruction without modifying the pre-trained model. The training optimises the following loss:

$$L = \gamma \cdot \text{MSE}(x_t, \hat{x}_t) + \text{CE}(d_t, v) \tag{2}$$

where $\text{MSE}(x_t, \hat{x}_t)$ minimises reconstruction error, and $\text{CE}(d_t, v)$ ensures accurate domain classification by the gate. This approach improves rate-distortion performance while preventing catastrophic forgetting of the source domain [40].

Unsupervised Domain Adaptation (UDA): UDA transfers knowledge from a labelled source domain to an unlabelled target domain, reducing the need for target labels [51]. It aligns feature distributions using adversarial training and feature transformation. This research employs DANN [15] and ADDA [48], two widely used UDA methods.

Domain-Adversarial Neural Network (DANN) integrates a Gradient Reversal Layer (GRL) to align feature distributions between source and target domains. It consists of shared feature extraction layers and two classifiers: one for label prediction and another for domain discrimination. The training objective minimises the label prediction loss while maximising domain confusion through GRL, enforc-

ing domain-invariant feature extraction. The loss function is:

$$L = L_y - \lambda L_d \tag{3}$$

where L_y is the label prediction loss, L_d is the domain classification loss, and λ controls the adversarial effect of GRL. This encourages the model to learn features that minimise classification error while preventing domain discrimination [15].

Adversarial Discriminative Domain Adaptation (ADDA) employs separate mappings for source and target domains with untied network weights. The target model is initialized from the pre-trained source model, enabling domain-specific feature extraction. ADDA aligns feature distributions by minimizing the distance between source and target embeddings. The objective function is:

$$L = L_s + \lambda D(F_s, F_t) \tag{4}$$

where L_s is the source classification loss, $D(F_s, F_t)$ is the discrepancy between source and target feature distributions, and λ controls adaptation strength. This iterative process refines target representations while preserving source knowledge [48].

4. Experiments

4.1. Implementation Details

Base Model: EfficientNet-B2, pretrained on ImageNet, is fine-tuned on a balanced, curated ISIC dermoscopic dataset. The trained model is evaluated on the IMPS dataset a combination of clinical images as well as on each individual subset (I, M, P, S).

SDA: For SDA, we implement ATDOC and LIC, using the balanced dermoscopic dataset as the source and labelled clinical images from IMPS(and its individual subsets) as the target. EfficientNet-B2 is used as the feature extractor in all cases.

UDA: For UDA, we employ DANN and ADDA, with EfficientNet-B2 as the shared feature extractor for both source and target. In these experiments, the source domain consists of the balanced dermoscopic dataset with labels, while the target domain comprises unlabelled clinical images from IMPS (and its individual subsets). DANN utilises a gradient reversal layer for domain-invariant feature learning, whereas ADDA trains a separate target feature extractor, initialised from the source model, to align the target representation.

Common Training Settings: All experiments are trained with a batch size of 16 and a learning rate of 1e-4. Each model is independently trained three times, and the final performance metrics are reported as the mean ± standard deviation across these runs.

4.2. Evaluation Metrics

For balanced dermoscopic datasets, accuracy is a reliable measure, but for imbalanced clinical datasets, AUROC, precision, recall, and F1 score provide a better assessment. AUROC accounts for class imbalance, precision prevents excessive false positives, and recall ensures critical melanoma cases are detected. F1 score balances these factors, making it suitable for imbalanced distributions. The average metric summarises overall performance, ensuring robust evaluation across cross-domain and domain adaptation settings.

4.3. Results and Discussion

This section evaluates the performance of our baseline model (EfficientNet-B2 trained on dermoscopic images) relative to the same architecture augmented with either supervised (ATDOC, LIC) or unsupervised (DANN, ADDA) domain adaptation. We also examine how the use of a more diverse target domain (IMPS) compares to evaluating on smaller, single-source subsets (I, M, P, S).

4.3.1. Supervised vs. Unsupervised Domain Adaptation Performance

Tables 3 and Table 4 summarise results for supervised domain adaptation (SDA) using ATDOC and LIC, respectively. Although these methods exhibit modest gains over the baseline on certain subsets, their improvements on the composite IMPS dataset remain constrained. For instance, ATDOC (Table 3) increases accuracy on IMPS from 58.55% (baseline in Table 2) to 68.55%, yet recall remains only 34.60% and F1 stands at 48.20%. Similarly, LIC (Table 4) yields 60.15% accuracy and an F1 of 48.18% on IMPS, which is better than the baseline's 41.25% F1 but still relatively modest for a supervised technique leveraging labelled target data. The subsets can offer more optimistic metrics—for example, LIC boosts accuracy on the MEDNODE subset to 81.40% yet even there, the gains are not always dramatic.

By contrast, unsupervised domain adaptation (UDA) with DANN (Table 5) achieves more pronounced improvements without requiring any target labels. On IMPS, DANN attains 68.10% accuracy, 63.45% F1, and 65.30% recall, comfortably surpassing both supervised approaches. Moreover, DANN significantly alleviates the baseline's difficulty in recognising melanoma on subsets such as PAD-UFES, boosting F1 to 74.32% from a baseline of 16.29%. ADDA (Table 6) shows sporadic successes such as higher recall on certain subsets but displays greater instability overall, particularly on PAD-UFES, where precision collapses to 5.20%. In general, DANN provides the most consistent benefits across multiple measures and domains, often outperforming the supervised methods. One reason why DANN

Table 2. Performance metrics for Base model (EfficientnetB2) on IMPS and (I,M,P,S) individually. Source is (Dermoscopic) and Target is (Clinical).

Setting	Dataset	AUROC	Accuracy	Precision	Recall	F1 Score	Average
Base	B→IMPS	88.60±0.02	58.55±0.01	72.05±0.02	29.40±0.04	41.25±0.01	57.97±0.01
	B→S	71.30±0.02	64.19±0.01	68.10±0.03	53.15±0.01	60.28±0.03	63.40±0.01
	B→I	62.37±0.03	47.24±0.05	78.30±0.05	24.18±0.01	37.44±0.08	49.91±0.02
	B→M	84.45±0.02	77.75±0.04	78.50±0.01	65.08±0.09	70.47±0.07	75.25±0.02
	B→P	77.34±0.02	90.26±0.02	10.36±0.05	37.10±0.04	16.29±0.01	46.27±0.01

Table 3. Performance metrics for SDA (ATDOC) on and (I, M, P, S) individually. Source: Dermoscopic, Target: Clinical.

Setting	Dataset	AUROC	Accuracy	Precision	Recall	F1 Score	Average
ATDOC (SDA)	B→IMPS	61.75±0.02	68.55±0.05	75.50±0.04	34.60±0.03	48.20±0.02	57.72±0.02
	B→S	63.82±0.06	60.55±0.04	62.10±0.02	48.36±0.03	54.28±0.02	57.82±0.02
	B→I	61.60±0.02	50.65±0.01	82.24±0.05	32.05±0.03	45.48±0.03	54.40±0.01
	B→M	85.35±0.02	77.68±0.01	74.46±0.02	71.25±0.03	72.15±0.01	76.18±0.01
	B→P	68.60±0.01	72.44±0.02	4.60±0.04	48.58±0.05	8.36±0.03	40.52±0.01

Table 4. Performance metrics for SDA (LIC) on and (I, M, P, S) individually. Source: Dermoscopic, Target: Clinical.

Setting	Dataset	AUROC	Accuracy	Precision	Recall	F1 Score	Average
LIC (SDA)	B→IMPS	70.65±0.05	60.15±0.03	73.08±0.03	32.29±0.02	48.18±0.01	56.87±0.01
	B→S	70.20±0.03	63.60±0.04	72.38±0.05	44.26±0.05	56.10±0.02	61.31±0.02
	B→I	71.10±0.02	48.04±0.03	85.03±0.01	22.45±0.04	35.41±0.06	52.41±0.02
	B→M	85.30±0.01	81.40±0.02	75.60±0.03	77.20±0.02	77.37±0.02	79.37±0.01
	B→P	68.77±0.07	82.30±0.02	7.58±0.04	46.60±0.03	12.37±0.05	43.52±0.02

may perform better is that it does not use target labels, so it may avoid overfitting to a small amount of labelled data. Potentially, it learns to match features between the two domains, which can help the model work better on different types of clinical images.

4.3.2. Performance on Diverse vs. Single-Target Domains

A key part of our contribution is evaluating under two target scenarios: a single-target domain (I, M, P, or S) versus the more diverse IMPS dataset. Observing the baseline (Table 2), we see that it achieves 77.75% accuracy and a 70.47% F1 on MED-NODE, which might suggest decent generalisability. However, once tested on IMPS, the same model's accuracy drops to 58.55% and F1 to 41.25%. This contrast underscores how using narrower target domains can mask real-world challenges. A model might appear effective on a single dataset but struggle under larger variations in lighting, imaging devices, and demographic factors.

Similarly, ATDOC and LIC can reach respectable results on certain subsets yet underperform on IMPS, revealing the difficulty of fully capturing the multi-source variability inherent in the combined domain. DANN's improvements on IMPS highlight its robustness in handling this additional diversity. Nonetheless, even DANN shows a dip in performance on IMPS compared to certain individual subsets, suggesting that each method's true stress test emerges only when evaluated on wide-ranging data.

4.3.3. Baseline vs. Adapted Models: Impact of Domain Shift

The disparity between baseline and adapted results confirms the substantial effect of domain shift between dermoscopic (source) and clinical (target) images. Table 2 shows how the baseline often yields high accuracy but low F1 and recall when tested on new clinical domains. On PAD-UFES, for instance, the baseline's accuracy is 90.26%, yet F1 is merely 16.29%, implying it is heavily biased toward the majority (benign) class. Both SDA (Tables 3 and Table 4) and UDA (Tables 5 and Table 6) mitigate this bias by aligning feature distributions in different ways. DANN stands out in its ability to remedy imbalanced predictions: on PAD-UFES, it raises F1 to 74.32% by improving recall and precision simultaneously.

This pattern repeats across other subsets. On ISIC (clinical) images (B→I), the baseline yields only 24.18% recall, whereas DANN raises it to 48.35%, and even ATDOC surpasses 30%. Although supervised approaches also yield noticeable improvements, the gains with DANN are typically larger. These findings suggest that unsupervised alignment

Table 5. Performance metrics for UDA (DANN) on and (I, M, P, S) individually. Source: Dermoscopic, Target: Clinical.

Setting	Dataset	AUROC	Accuracy	Precision	Recall	F1 Score	Average
DANN (UDA)	B→IMPS	60.40±0.04	68.10±0.02	60.33±0.04	65.30±0.04	63.45±0.10	63.52±0.02
	B→S	66.20±0.01	61.30±0.02	58.40±0.02	72.25±0.11	66.25±0.12	64.88±0.03
	B→I	66.20±0.05	56.38±0.09	78.40±0.02	48.35±0.01	61.22±0.05	62.11±0.02
	B→M	87.30±0.02	79.40±0.03	72.10±0.02	79.16±0.01	74.65±0.02	78.52±0.01
	B→P	53.60±0.06	59.30±0.01	95.20±0.08	60.28±0.07	74.32±0.06	68.54±0.03

Table 6. Performance metrics for UDA (ADDA) on and (I, M, P, S) individually. Source: Dermoscopic, Target: Clinical.

Setting	Dataset	AUROC	Accuracy	Precision	Recall	F1 Score	Average
ADDA (UDA)	B→IMPS	46.40±0.09	50.40±0.04	50.10±0.02	80.05±0.11	65.33±0.07	58.46±0.03
	B→S	49.70±0.03	48.47±0.08	50.25±0.06	80.10±0.46	67.40±0.04	59.18±0.09
	B→I	67.28±0.05	57.15±0.07	65.24±0.08	95.70±0.02	79.12±0.03	72.90±0.02
	B→M	42.36±0.07	38.50±0.10	42.38±0.05	1.00±0.02	58.39±0.08	36.53±0.03
	B→P	25.56±0.40	48.70±0.37	5.20±0.21	90.45±0.33	0.07±0.41	34.00±0.16

can be particularly powerful when the target domain's distribution is significantly different or insufficiently captured by the limited labelled target data in SDA.

4.3.4. Overestimation of Robustness in Narrow Evaluations

Finally, the results show that restricting evaluation to a single or less varied target domain can overestimate model robustness. The baseline's performance on MED-NODE alone appears moderately successful, but this success does not transfer to the multi-faceted IMPS dataset. Likewise, its ostensibly high accuracy on PAD-UFES conceals an inability to detect melanoma. Had this paper confined itself to only one or two target sets, it might have overlooked these shortcomings. Instead, combining multiple clinical image sources into the IMPS dataset reveals the model's real fragility and the importance of applying domain adaptation methods capable of handling broad distributional shifts. Hence, models validated solely on limited data may be inadequately tested for real-world conditions, and the results here underscore the need for diverse clinical evaluations to avoid inflated expectations of robustness.

5. Limitations

While our results demonstrate that domain adaptation can substantially improve generalisation when clinical images are more diverse than the training domain, our conclusions would be more robust if we could evaluate the models on an entirely separate set of unseen clinical images not used for any phase of training. Such a cross-domain test could involve both less diverse and more extensively varied distributions, enabling a finer-grained analysis of model robustness under different clinical settings. However, the limited public availability of large-scale, heterogeneous clinical skin cancer datasets currently hinders more exhaustive experimentation, leaving an important avenue for future work. All experiments were done using EfficientNet-B2 due to its effectiveness and efficiency in image classification tasks. Experiments with other models will be explored in the future.

6. Conclusion

This study demonstrates the importance of addressing domain shifts in skin cancer classification. Our findings show that unsupervised domain adaptation, particularly DANN, improves generalisation more effectively than supervised methods. The diverse IMPS dataset exposed model weaknesses that narrower evaluations did not capture, emphasising the need for comprehensive testing. While our results confirm the benefits of domain adaptation, the study is limited by dataset availability. Future research should evaluate these methods on additional clinical datasets that were not part of this study to better assess their applicability in real-world settings.

References

[1] Ilán Carretero, Pablo Meseguer, Rocío del Amor, and Valery Naranjo. Enhancing whole slide image classification through supervised contrastive domain adaptation. 2024. 2

[2] Bill Cassidy, Connah Kendrick, Andrzej Brodzicki, Joanna Jaworek-Korjakowska, and Moi Hoon Yap. Analysis of the ISIC image datasets: Usage, benchmarks and recommendations. *Medical image analysis*, 75:102305, 2022. 3

[3] Sireesha Chamarthi, Katharina Fogelberg, Roman C. Maron, Titus J. Brinker, and Julia Niebling. Mitigating the influence of domain shift in skin lesion classification: A benchmark study of unsupervised domain adaptation methods on dermoscopic images. 2023. 3

[4] Sireesha Chamarthi, Katharina Fogelberg, Roman C Maron, Titus J Brinker, and Julia Niebling. Mitigating the influence of domain shift in skin lesion classification: A benchmark

study of unsupervised domain adaptation methods on dermoscopic images. *arXiv preprint arXiv:2310.03432*, 2023. 1

[5] Sireesha Chamarthi, Katharina Fogelberg, Titus J Brinker, and Julia Niebling. Mitigating the influence of domain shift in skin lesion classification: A benchmark study of unsupervised domain adaptation methods. *Informatics in Medicine Unlocked*, 44:101430, 2024. 1

[6] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng-Ann Heng. Aaai-19) provincial key laboratory of computer vision and virtual reality technology, shenzhen institutes of advanced technology. In *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, 2019. 3

[7] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. 3

[8] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018. 3

[9] Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, et al. Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*, 2019. 3

[10] Gabriela Csurka. Domain adaptation for visual applications: A comprehensive survey. *arXiv preprint arXiv:1702.05374*, 2017. 5

[11] Thomas de Bel, Meyke Hermsen, Rumcnl Jesper Kers, Jeroen van der Laak, and Geert Litjens. Stain-Transforming Cycle-Consistent Generative Adversarial Networks for Improved Segmentation of Renal Histopathology. In *Proceedings of Machine Learning Research*, pages 151–163, 2019. 3

[12] Zvi Drezner, Ofir Turel, and Dawit Zerom. A modified kolmogorov–smirnov test for normality. *Communications in Statistics—Simulation and Computation®*, 39(4):693–704, 2010. 4

[13] Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R Arabnia. A brief review of domain adaptation. *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pages 877–894, 2021. 1

[14] Katharina Fogelberg, Sireesha Chamarthi, Roman C Maron, Julia Niebling, and Titus J Brinker. Domain shifts in dermoscopic skin cancer datasets: Evaluation of essential limitations for clinical translation. *New Biotechnology*, 76:106–117, 2023. 1, 2, 3

[15] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016. 3, 5, 6

[16] Ioannis Giotis, Nynke Molders, Sander Land, Michael Biehl, Marcel F. Jonkman, and Nicolai Petkov. MED-NODE: A computer-assisted melanoma diagnosis system using non-dermoscopic images. *Expert systems with applications*, 42 (19):6578–6585, 2015. 2, 4

[17] Michael Goetz, Christian Weber, Franciszek Binczyk, Joanna Polanska, Rafal Tarnawski, Barbara Bobek-Billewicz, Ullrich Koethe, Jens Kleesiek, Bram Stieltjes, and Klaus H. Maier-Hein. Dalsa: Domain adaptation for supervised learning from sparsely annotated mr images. *IEEE Transactions on Medical Imaging*, 35:184–196, 2016. 2

[18] Manu Goyal, Thomas Knackstedt, Shaofeng Yan, and Saeed Hassanpour. Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities. *Computers in biology and medicine*, 127: 104065, 2020. 1

[19] Vanessa Gray-Schopfer, Claudia Wellbrock, and Richard Marais. Melanoma biology and new targeted therapy. *Nature*, 445(7130):851–857, 2007. 1

[20] Yanyang Gu, Zongyuan Ge, C Paul Bonnington, and Jun Zhou. Progressive transfer learning and adversarial domain adaptation for cross-domain skin disease classification. *IEEE journal of biomedical and health informatics*, 24(5):1379–1393, 2019. 2

[21] Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021. 1

[22] Hao Guan and Mingxia Liu. Domain Adaptation for Medical Image Analysis: A Survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2022. 3

[23] David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*, 2016. 3

[24] Johan Fredin Haslum, Christos Matsoukas, Karl Johan Leuchowius, and Kevin Smith. Bridging generalization gaps in high content imaging through online self-supervised domain adaptation. In *Proceedings - 2024 IEEE Winter Conference on Applications of Computer Vision, WACV 2024*, pages 7723–7732. Institute of Electrical and Electronics Engineers Inc., 2024. 3

[25] Lukas Hedegaard, Omar Ali Sheikh-Omar, and Alexandros Iosifidis. Supervised domain adaptation: A graph embedding perspective and a rectified experimental protocol. 2020. 2

[26] Qingshan Hou, Yaqi Wang, Peng Cao, Shuai Cheng, Linqi Lan, Jinzhu Yang, Xiaoli Liu, and Osmar R. Zaiane. A collaborative self-supervised domain adaptation for low-quality medical image enhancement. *IEEE Transactions on Medical Imaging*, 43:2479–2494, 2024. 3

[27] Zhe Huang, Ruijie Jiang, Shuchin Aeron, and Michael C Hughes. Systematic comparison of semi-supervised and self-supervised learning for medical image classification. In *Computer Vision and Pattern Recognition*. IEEE, 2024. 2

[28] ISIC. ISIC Archive. 2, 4

[29] K Kanchana, S Kavitha, KJ Anoop, and B Chinthamani. Enhancing skin cancer classification using efficient net b0-b7 through convolutional neural networks and transfer learning with patient-specific data. *Asian Pacific Journal of Cancer Prevention: APJCP*, 25(5):1795, 2024. 4

[30] Suruchi Kumari and Pravendra Singh. Deep learning for unsupervised domain adaptation in medical imaging: Recent advancements and future perspectives. *Computers in Biology and Medicine*, 170, 2024. 3

[31] Rui Li, Qianfen Jiao, Wenming Cao, Hau San Wong, and Si Wu. Model Adaptation: Unsupervised Domain Adaptation without Source Data. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 9638–9647. IEEE Computer Society, 2020. 3

[32] Shuang Li, Chi Harold Liu, Qiuxia Lin, Binhui Xie, Zhengming Ding, Gao Huang, and Jian Tang. Domain conditioned adaptation network. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, 2020. 3

[33] Jian Liang, Dapeng Hu, and Jiashi Feng. Domain adaptation with auxiliary target domain-oriented classifier. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16632–16642, 2021. 5

[34] Jia Liu, Wenjie Xuan, Yuhang Gan, Yibing Zhan, Juhua Liu, and Bo Du. An end-to-end supervised domain adaptation framework for cross-domain change detection. *Pattern Recognition*, 132, 2022. 2

[35] Roman C Maron, Justin G Schlager, Sarah Haggenmüller, Christof von Kalle, Jochen S Utikal, Friedegund Meier, Frank F Gellrich, Sarah Hobelsberger, Axel Hauschild, Lars French, et al. A benchmark for neural network robustness in skin cancer classification. *European Journal of Cancer*, 155: 191–199, 2021. 1

[36] Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. 2017. 2

[37] Abbas Omidi, Aida Mohammadshahi, Neha Gianchandani, Regan King, Lara Leijser, and Roberto Souza. Unsupervised domain adaptation of mri skull-stripping trained on adult data to newborns. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7718–7727, 2024. 3

[38] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019. 1

[39] Andre G. C. Pacheco, Gustavo R. Lima, Amanda S. Salomão, Breno A. Krohling, Igor P. Biral, Gabriel G. De Angelo, Fábio C. R. Alves, José G. M. Esgario, Alana C. Simora, Pedro B. C. Castro, Felipe B. Rodrigues, Patricia H. L. Frasson, Renato A. Krohling, Helder Knidel, Maria C. S. Santos, Rachel B. Do Espírito Santo, Telma L. S. G. Macedo, Tania R. P. Canuto, and Luíz F. S. De Barros. PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in brief*, 32:106221, 2020. 2, 4

[40] Alberto Presta, Gabriele Spadaro, Enzo Tartaglione, Attilio Fiandrotti, and Marco Grangetto. Domain adaptation for learned image compression with supervised adapters. In *2024 Data Compression Conference (DCC)*, pages 33–42. IEEE, 2024. 5

[41] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific data*, 8(1):34, 2021. 3

[42] Yaxuan Song, Jianan Fan, Dongnan Liu, and Weidong Cai. Multi-source-free domain adaptation via uncertainty-aware adaptive distillation. 2024. 3

[43] Karin Stacke, Gabriel Eilertsen, Jonas Unger, and Claes Lundström. Measuring domain shift for deep learning in histopathology. *IEEE journal of biomedical and health informatics*, 25(2):325–336, 2020. 1

[44] Baochen Sun and Kate Saenko. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. In *Computer Vision – ECCV 2016 Workshops*, pages 443–450, Amsterdam, The Netherlands, 2016. 3

[45] Xiaoxiao Sun, Jufeng Yang, Ming Sun, and Kai Wang. *A benchmark for automatic visual classification of clinical skin disease images*. 2016. 1, 2, 4

[46] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4068 – 4076, 2015. 2

[47] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pages 2962–2971. Institute of Electrical and Electronics Engineers Inc., 2017. 3

[48] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 1, 5, 6

[49] Janet Wang, Yunbei Zhang, Zhengming Ding, and Jihun Hamm. Achieving Reliable and Fair Skin Lesion Diagnosis via Unsupervised Domain Adaptation. In *Computer Vision and Pattern Recognition Workshop*, 2023. 3

[50] Janet Wang, Yunbei Zhang, Zhengming Ding, and Jihun Hamm. Achieving reliable and fair skin lesion diagnosis via unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5157–5166, 2024. 2

[51] Garrett Wilson and Diane J Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46, 2020. 5

[52] Yinhao Wu, Bin Chen, An Zeng, Dan Pan, Ruixuan Wang, and Shen Zhao. Skin cancer classification with deep learn-

ing: a systematic review. *Frontiers in Oncology*, 12:893972, 2022. 1

[53] Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning Semantic Representations for Unsupervised Domain Adaptation. In *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Sweden, 2018. 3

[54] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014. 1

[55] Lei Zhang, Guang Yang, and Xujiong Ye. Automatic skin lesion segmentation by coupling deep fully convolutional networks and shallow network with textons. *Journal of Medical Imaging*, 6(2):024001–024001, 2019. 1