Achieving Reliable and Fair Skin Lesion Diagnosis via Unsupervised Domain Adaptation

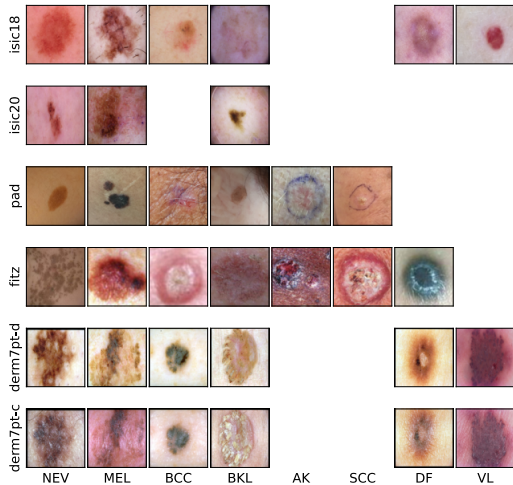## Supplementary Material

Figure 5. Image examples of each dataset considered in this study.
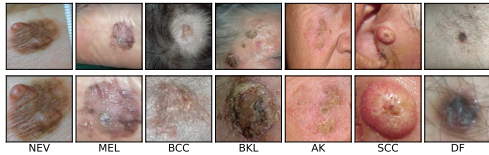

Figure 6. Examples of ROI-cropped vs. uncropped images for Fitzpatrick17k.

6. Image Examples of Selected Datasets

Fig. 5 shows example images for each class and dataset considered in this study. Images from the Fitzpatrick17k dataset have been processed with ROI-based cropping. It is clear that images from different datasets are visually distinct, even for the same conditions, leading to a significant domain gap and posing challenges for classification.

7. ROI Pre-processing

In [13], region of interest (ROI) detection is utilized to separate skin lesions from clinical photos, effectively reducing noise and enhancing the lesion information ratio. In the context of our problem setting, dermoscopic images of skin lesions are mostly close-up shots centered on the lesions, whereas clinical photos are taken at varying distances from the lesions or from different angles. To minimize background noise, as well as the discrepancy between dermoscopic and clinical images, we fine-tune a YOLO-8 model, one of the SOTA ROI detection algorithms [16]. After cropping, each image is resized to a resolution of 64×64 pixels, which is the designated image size for subsequent experiments. Fig. 6 are examples of ROI-cropped images.

Domain	Metric	Single	Single DANN	Combined	Combined DANN	M³SDA
fitz (skin color)	PQD	$82.4\pm_{1.2}$	$74.1\pm_{1.9}$	$79.6\pm_{1.0}$	$\mathbf{87.1\pm_{1.8}}$	$83.2\pm_{0.6}$
	DPM	$56.7\pm_{1.7}$	$81.7\pm_{0.7}$	$92.4\pm_{1.4}$	$\mathbf{93.5\pm_{1.6}}$	$90.8\pm_{1.1}$
	EOM	$69.7\pm_{1.7}$	$75.9\pm_{1.9}$	$74.8\pm_{0.8}$	$\mathbf{78.4\pm_{1.4}}$	$77.7\pm_{1.4}$
	AUROC	$69.9\pm_{0.9}$	$84.3\pm_{3.3}$	$84.0\pm_{2.5}$	$\mathbf{92.2\pm_{5.7}}$	$83.4\pm_{3.2}$
isic2020 (age)	PQD	$94.1\pm_{1.0}$	$95.8\pm_{1.0}$	$98.7\pm_{1.0}$	$99.0\pm_{0.8}$	$\mathbf{99.1\pm_{0.4}}$
	DPM	$77.9\pm_{1.0}$	$82.0\pm_{1.4}$	$92.6\pm_{0.6}$	$\mathbf{98.0\pm_{3.0}}$	$89.4\pm_{1.7}$
	EOM	$79.6\pm_{1.2}$	$81.3\pm_{2.3}$	$88.3\pm_{2.4}$	$92.9\pm_{1.4}$	$\mathbf{95.4\pm_{2.9}}$
	AUROC	$70.9\pm_{2.3}$	$71.4\pm_{3.0}$	$87.8\pm_{3.9}$	$87.3\pm_{3.3}$	$\mathbf{89.4\pm_{3.9}}$
pad (age)	PQD	$83.7\pm_{2.2}$	$86.0\pm_{2.0}$	$87.1\pm_{1.1}$	$88.5\pm_{1.4}$	$\mathbf{96.1\pm_{2.5}}$
	DPM	$75.4\pm_{0.4}$	$81.7\pm_{1.7}$	$73.1\pm_{1.1}$	$\mathbf{95.2\pm_{2.3}}$	$88.8\pm_{2.2}$
	EOM	$54.7\pm_{1.7}$	$47.3\pm_{2.3}$	$46.3\pm_{3.3}$	$48.4\pm_{1.4}$	$\mathbf{68.9\pm_{1.4}}$
	AUROC	$65.3\pm_{2.3}$	$74.0\pm_{3.0}$	$74.9\pm_{3.9}$	$76.9\pm_{3.3}$	$\mathbf{84.9\pm_{3.9}}$

Table 5. Fairness Evaluation Metrics across Different Domains

8. Fairness Evaluation

Fig. 7 shows the distribution of sensitive attributes across each dataset considered in the fairness evaluation. In the Fitzpatrick17k dataset, skin color is identified as a sensitive attribute. Notably, the dataset is skewed towards the light-skinned sub-population (FST 1-2), making images with dark skin tone (FST 5-6) a minority group. This imbalance may lead to under-diagnosis in dark-skinned individuals when using AI for early screening. Similarly, for ISIC2020 and PAD-UFES-20, age is considered a sensitive attribute, dividing the data into two groups. These datasets tend to be skewed towards individuals older than 30, potentially resulting in a higher risk of under-diagnosis in the younger population during AI-based early screening.
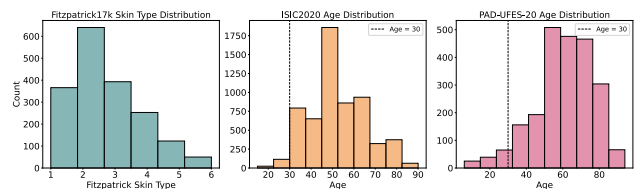

Figure 7. Sensitive Attribution Distribution for Datasets in Fairness Evaluation.