Supplementary Materials for FEDD - Fair, Efficient, and Diverse Diffusion-based Lesion Segmentation and Malignancy Classification

Héctor Carrión[1*], Narges Norouzi[2]

University of California, Santa Cruz[1], University of California, Berkeley[2],
hcarrion@ucsc.edu*

1 Dataset and annotation details

As mentioned on the main text, we draw 4 balanced subsets of DDI for training, each representing approximately 5% (10 samples per skin tone), 10% (20 samples per skin tone), 15% (30 samples per skin tone), and 20% (40 samples per skin tone) of DDI. The smaller training sets are subsets of the larger ones, this is to say 5% $\subseteq$ 10% $\subseteq$ 15% $\subseteq$ 20% $\subseteq$ DDI. For classification we draw validation and test sets, each containing 30 samples (10 samples per skin tone). Further, we test model checkpoints trained on each DDI subset on all remaining DDI images (476 samples), accuracy results from this larger test set are reported on the paper text and on supplementary materials Table 3. For segmentation we test on 198 additionally annotated samples (59 light, 80 medium and 59 dark in skin-tone). These larger test sets are skin-tone unbalanced, as expected in clinical settings. All validation and testing sets are disjoint from the training splits and from each other.

For malignancy classification, DDI includes disease labels. For segmentation, sample images need to be semantically labeled and some may be difficult to annotate; the annotation protocol is as follows:

1. Lesion is segmented following the boundary at which the skin transitions from healthy to unhealthy appearance.
2. Markings or rulers are segmented in their visible totality.
3. Non-lesion, or normal skin is segmented.

This denoted our segmentation masks which cover 5 different classes: lesion, skin, marker, ruler, and background. We opted to label these classes as many images include a ruler and markings to denote the lesion of focus. All other potentially present objects, like clothing, were denoted as background.

Not all images were suitable for labeling as DDI images may not have been collected with computer vision in mind and consequently were skipped. The following criteria will trigger a skip:

1. Lesion is occluded.
2. Lesion is significantly blurry.

3. Lesion is only partially visible.
4. Target lesion is ambiguous (no lesions marked when multiple are present or multiple are marked in a single image).
5. Lesion is on or near the scalp (hair is not a labeled target and time consuming to annotate).

Examples of images which triggered a skip can be found on DDI sample id 25, 55, and 161. The total number of annotated images which passed our annotation criteria and secondary review is 378. We have released this annotation work on our github linked in the abstract. Table 1 describes our training subsets of the Diverse Dermatology Images dataset.

Table 1. The distribution of training data per sub-set of DDI.

DDI Subset	Total Samples	Samples per Skin Tone	Malignant Samples per Skin Tone	Benign Samples per Skin Tone
5%	30	10	5	5
10%	60	20	10	10
15%	90	30	15	15
20%	120	40	20	20

2 Software and Hardware

2.1 Setup

All code involving this study was written in Python v3.7.12. The packages used for training and inference were the latest available stable versions of deep learning frameworks PyTorch v1.13, Keras v2.8, and TensorFlow v2.8. Additional packages used for numerical processing, plotting and visualizations were Pandas v1.3.5, NumPy v1.21.5, sklearn v1.0.2, PIL 7.1.2, and Matplotlib 3.5.1. The hardware used was a Google Colab instance running a quad-core Intel Xeon CPU at 2.3 GHz with 80GB of RAM and an NVIDIA A100 GPU with 40GB of vRAM.

2.2 Performance

The average run-time for FEDD training was about 40 minutes per MLP over 100 epochs. Memory footprint did not exceed 80GB of system memory or 40GB GPU memory after multiple runs. Inference time is equal to about 1.5 images per second.

3 Additional Hyper-parameters and Experimental Setup

Additional hyper-parameters used were 30 batch size, 60 batch size, 90 batch size, and 120 batch size for 5, 10, 15, and 20% of DDI, respectively. These were selected as they allowed all subset data to be processed in a single batch. For testing and validation, a 30-batch size was used. The Adam optimizer was used, along with weight decay equal to 0.00001. The loss function is Binary Cross Entropy. The MLPs were trained over 500 epochs with early stopping monitoring validation loss. Most runs early-stopped at or before 100 epochs.

4 K-Means Clustering

Figure 1 represents some examples of clustering quality for FEDD block activations and embeddings. K-Means was used with $K = 3$.

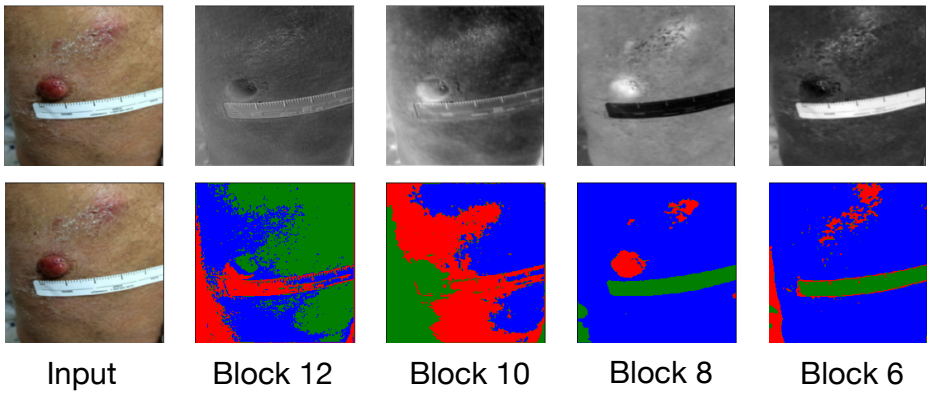

Fig. 1. FEDD activation maps (top) and their clustering quality (bottom).

5 ROC Curves

Figure 2 showcases the ROC curves for previous state-of-the-art and FEDD.

6 F1 Performance per Skin Tone

Table 2 demonstrates the F1 performance per skin tone for FEDD and all other tested architectures.

7 Improvement Deltas

Table 3 demonstrates the performance advantage of FEDD versus the next best tested method.

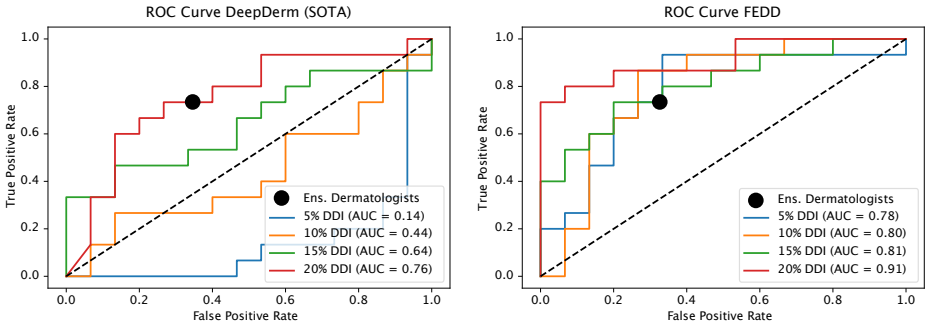

Fig. 2. Test set ROC Curve for DeepDerm (left) and FEDD framework (right) at each subset of DDI data. The performance of an ensemble of dermatologists is also shown as a black circle.

Table 2. F1 Performance for all tested architectures split between light (left) and dark (right) skin tones.

Method	5% DDI (Light)	10% DDI (Light)	15% DDI (Light)	20% DDI (Light)	5% DDI (Dark)	10% DDI (Dark)	15% DDI (Dark)	20% DDI (Dark)
DenseNet121	0.60	0.62	0.57	0.67	**0.80**	0.71	0.57	0.77
VGG16	0.62	0.50	**0.73**	0.73	0.67	0.57	0.62	0.77
ResNet50	0.00	0.25	0.29	0.00	0.33	0.60	0.50	0.00
EfficientNetB0	0.67	0.00	0.29	0.62	0.46	0.00	0.44	**0.91**
MobileNetV2	0.55	0.33	0.62	**0.89**	0.67	0.75	0.55	0.89
DeepDerm	0.67	0.00	0.44	0.60	0.67	0.50	0.57	0.75
FEDD	**0.73**	**0.83**	0.67	**0.89**	**0.80**	**0.83**	**0.75**	0.89

Table 3. Performance Improvement of FEDD compared to next-best method on the full segmentation and classification test sets (skin-tone unbalanced).

Metric	5% DDI	10% DDI	15% DDI	20% DDI
IoU	0.18	0.13	0.06	0.07
Accuracy	14%	6%	5%	4%

8 Expanded results visualization

Fig 3 showcases additional segmentation results in full-width for easy viewing.

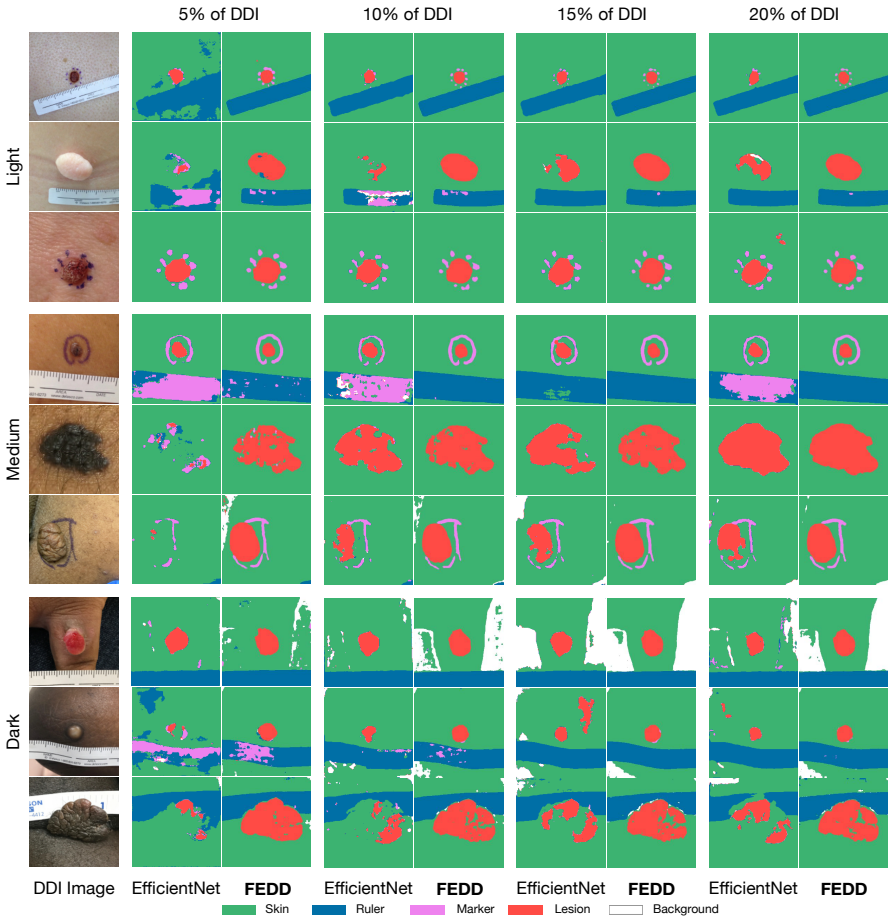

Fig. 3. Expanded test set segmentation results for FEDD and EfficientNet.