Supplementary Materials

Healthc Inform Res. 2022 July;28(3):222-230.
https://doi.org/10.4258/hir.2022.28.3.222
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Table S1. Dermatologists' disease severity grading (A) distribution

Diagnosis	S0	S1	S2	S3	S4	Total
Acropustulosis of infancy	1	9	5	2	0	17
Palmoplantar pustular psoriasis	11	20	21	39	4	95
Pustulosis palmoplantaris	0	27	20	14	0	61
Pustulosis subcornealis	0	9	19	12	0	40
All diagnoses	12	65	65	67	4	213

Table S2. Medical student's lesion count ranking (B) distribution

Diagnosis	S0	S1	S2	S3	S4	Total
Acropustulosis of infancy	2	7	6	1	1	17
Palmoplantar pustular psoriasis	10	35	21	21	8	95
Pustulosis palmoplantaris	0	27	21	8	5	61
Pustulosis subcornealis	0	20	6	10	4	40
All diagnoses	12	89	54	40	18	213

Table S3. Correlation coefficients of predictions aggregated on full images

	ICC	
	Surface	Count
Pustules	0.94 (0.89–0.97)	0.99 (0.98–1.00)
Brown spots	0.96 (0.93–0.99)	0.98 (0.98–0.99)
All lesions	0.98 (0.96–0.99)	0.99 (0.98–1.00)

The values in parentheses correspond to the 95% confidence interval. Results obtained from the 30 images in the test set.

ICC: intraclass correlation coefficient.

All *p*-values are below 0.05.

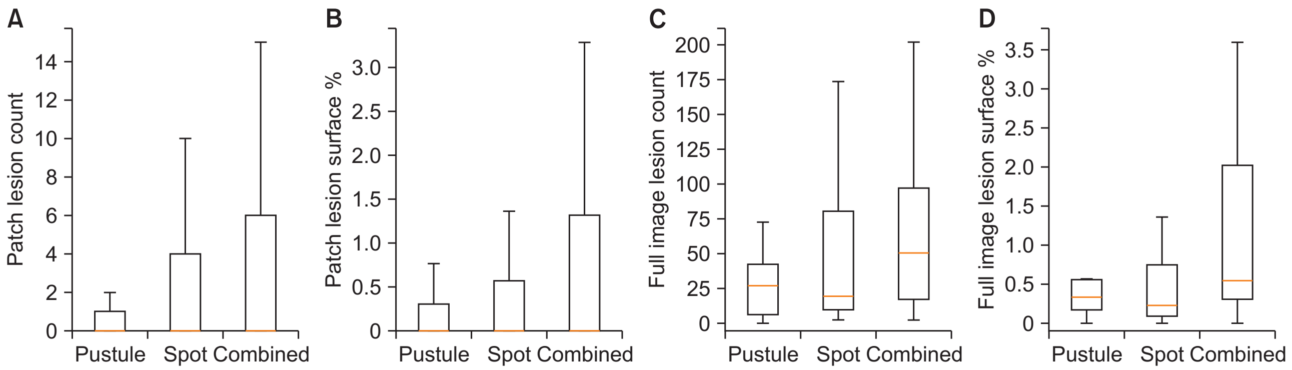

Figure S1. PPP test set lesion distribution. Plots (A) and (B) show, respectively, the count and surface distribution for image patches in the test set. Plots (C) and (D) show the same for the corresponding full images. PPP: palmoplantar pustular psoriasis.

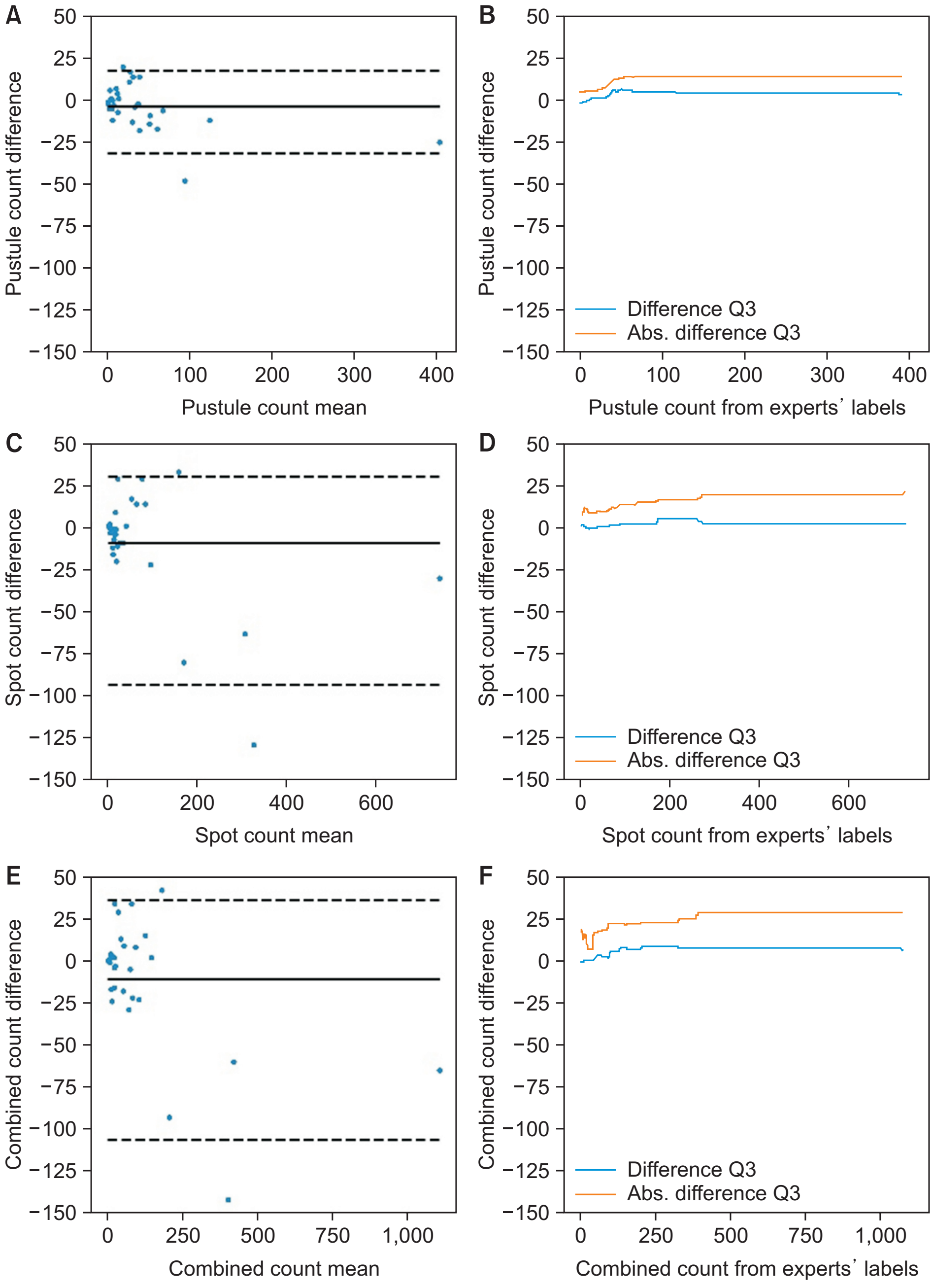

Figure S2. Agreement of count predictions with expert labels on full images. The DLM predictions differed by at most 22.5 lesions in 75% of the patches with up to 97 lesions (the test set's Q3). For the remaining patches, the difference increased to 29 lesions in 75% of the cases. The DLM's bias was -11.1 for both types of lesions, its MAD was 23.96, and the ICC was 0.99 (95% CI, 0.98–1.00). DLM: deep learning model, MAD: mean absolute difference, ICC: intraclass correlation coefficient, CI: confidence interval.

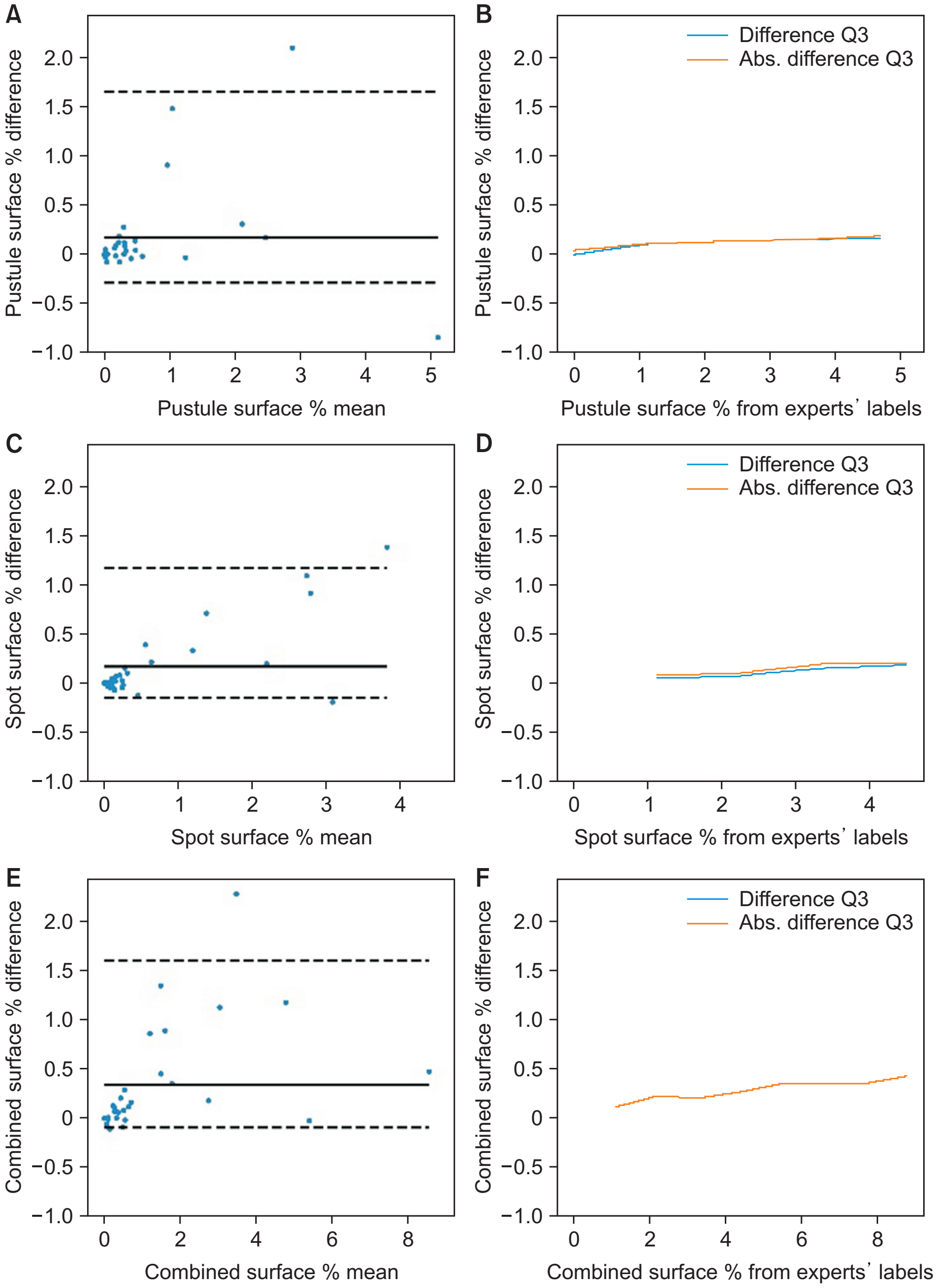

Figure S3. Agreement of surface predictions with expert labels on full images. Considering the test image patches with up to 2% (the test set Q3) of the skin surface covered by pustules and brown spots, the DLM was able to determine the surface with less than 0.22% difference from dermatologists in 75% of the cases. This difference plateaued at 0.42% for 75% of the images with higher surface percentages. The predicted surface ratios of lesions related to the experts' labels with an ICC of 0.98 (95% CI, 0.96–0.99). The DLM bias was 0.33% while the MAD was 0.35%. DLM: deep learning model, MAD: mean absolute difference, ICC: intraclass correlation coefficient, CI: confidence interval.

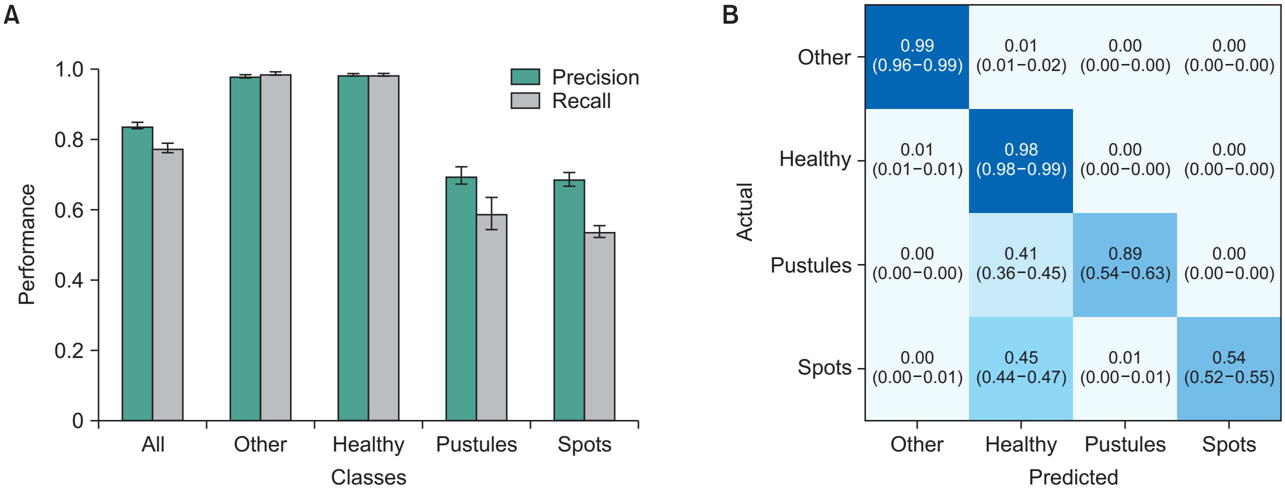

Figure S4. Pixel-wise performance of the DLM in segmentation. Plot (A) shows the pixel precision and recall reached on the test set by the DLM. The first two bars, for the “all” category, represent the macro average of the classes’ individual performance. Plot (B) is a confusion matrix showing the mean proportion of pixels classified among the different classes. Its vertical axis represents the true pixel labels, while the horizontal axis shows the predicted labels. The error bars and values in parentheses represent the 95% confidence interval. The evaluation of the DLM’s pixel-wise performance showed a precision and a recall of 69% and 59% respectively for pustules, and 68% and 54% for brown spots. The DLM missed 41% of pustules pixels and 45% of brown spots pixels, matching the previous observation that it underestimated the lesion sizes. These relatively low scores are a direct consequence of the idiosyncrasy of the experts’ labels. We also evaluated the segmentation performance without ImageNet pretraining and observed a drop in performance. For pustules, we calculated a precision of 35% and recall of 36%, while for brown spots the precision was 48% and the recall was 47%. According to the DLM hyperparameters, with cross-validation, we selected the following hyperparameters for both skin and lesion segmentation DLMs: the batch size was 16, the initial learning rate was 1e-4, the input size was 380 × 380 pixels, and the number of epochs was 40. DLM: deep learning model.

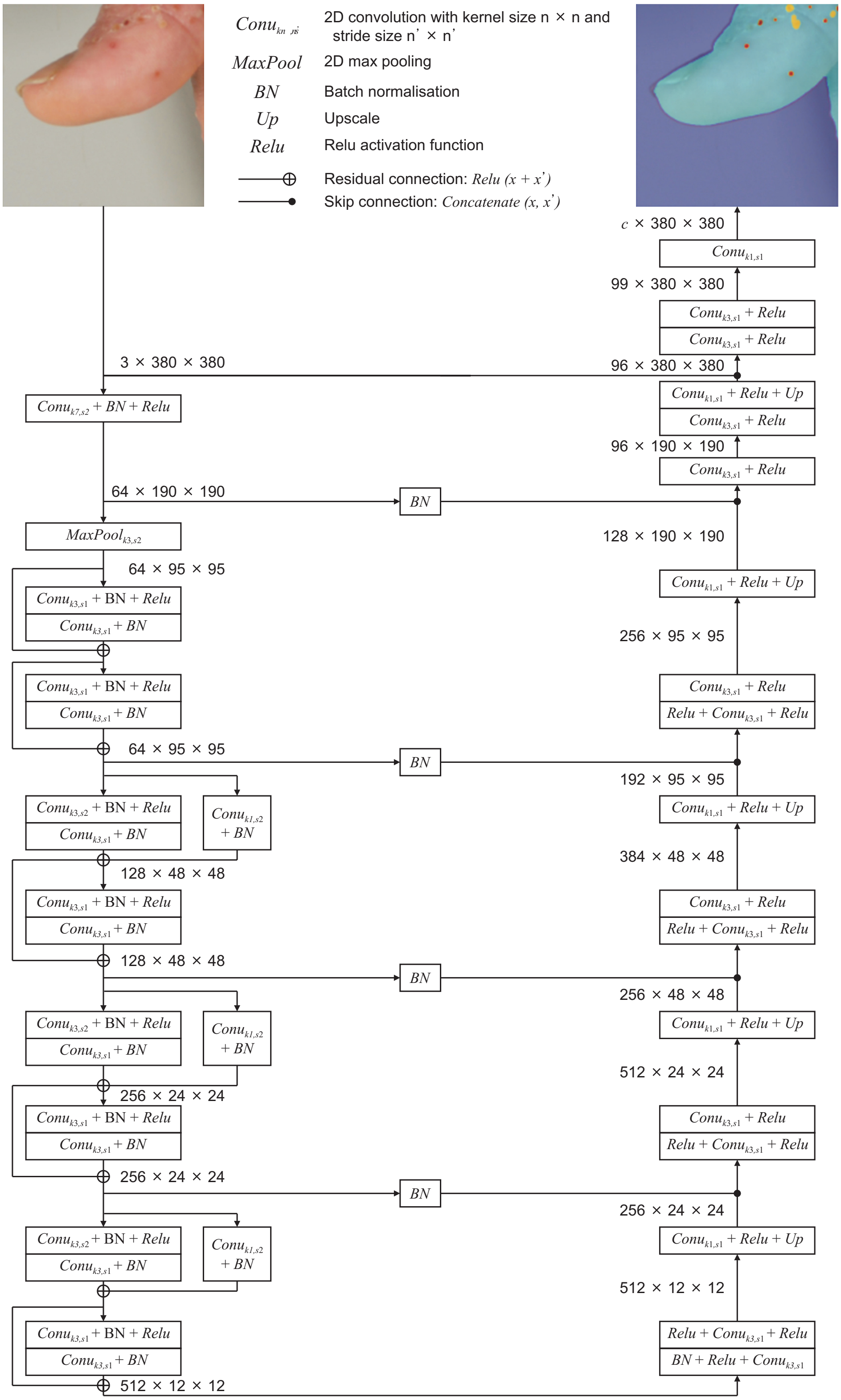

Figure S5. Architecture of the deep learning segmentation model. This figure presents the structure of the segmentation models, based on the U-Net and ResNet architectures. The final mask channel dimension was c = 2 for skin segmentation (M1 subunit) and c = 3 for lesion segmentation (M2 subunit).