Deep Learning Dichotomous Classification of Inflammatory Changes in Keratinocyte Carcinomas Treated with Topical Therapy with Imiquimod, 5-fluorouracil, and Retinoids

William J. Nahm, BA[1]; Olivia L. Shen, BS[2]; Keyvan Nouri, MD, MBA[3,4]; Robert S. Kirsner, MD, PhD[3,4]; George W. Elgart, MD[3]; John Tsatalis, MD[3]; Claire L. Shen, BS[5]; Flor D. Valadez, CMA[5]; Anthony Wu, BS[6]; Evangelos V. Badiavas, MD, PhD[3]

1: New York University Grossman School of Medicine, New York, NY, USA; 2: University of California, Davis, Davis, CA, USA; 3: Dr. Phillip Frost Department of Dermatology & Cutaneous Surgery, University of Miami Miller School of Medicine, Miami, FL, USA; 4: Sylvester Comprehensive Cancer Center, Miami, FL, USA; 5: Shen Dermatology, Temecula, CA, USA; 6: University of California, San Diego, San Diego, CA, USA

UHealth
UNIVERSITY OF MIAMI HEALTH SYSTEM

INTRODUCTION

•Topical combination therapy with imiquimod, 5-fluorouracil, and retinoids (IMI/5-FU/RET) provides a cost-effective treatment option for keratinocyte carcinomas (KCs) that facilitates remote patient monitoring.
•Clinical response monitoring includes assessment of inflammation intensity, which reflects treatment activity and guides therapy duration.
•Artificial intelligence (AI)-based inflammation classification could reduce in-person visits while maintaining monitoring quality.
•Deep learning (DL) has demonstrated dermatologist-level performance in classifying lesions; however, application to inflammation assessment in KC treatment remains unexplored.
•This study evaluates three DL approaches for binary classification of "active" versus "inactive/minimal" inflammation status in KCs treated with IMI/5-FU/RET.

METHODS

•The SC-2000 dataset is the first set of images involved in KC treatment inflammation classification, comprising 2,170 RGB images from 577 patients covering 861 KCs (~65% basal cell carcinomas [BCCs], ~35% squamous cell carcinomas [SCCs]) captured before, during (weeks 1-16), and after IMI/5-FU/RET treatment (2011-2024).
•Images were labeled as "**active**" (intense erythema, scaling, oozing) or "**inactive/minimal**" (mild residual erythema or post-inflammatory changes) (**Figure 1**) by the physician.
•The labeled dataset was split into training (70%), validation (20%), and testing (10%) while preventing patient-level data leakage (**Table 1**).
•Images underwent preprocessing and augmentation (**Figure 2**) to evaluate three DL approaches: (1) custom convolutional neural network (CNN), (2) pre-trained feature extractors (ResNet50, MobileNetV2, DenseNet121) with linear probing, and (3) pre-trained feature extractors with K-nearest neighbor (K-NN) similarity-based classification.
•Test set performance was evaluated using accuracy, recall, precision, and F1 score with 95% confidence intervals calculated via patient-level bootstrapping (5,000 iterations).

Table 1. Distribution of Images, Patients, and Lesions Across Training, Validation, and Test Sets Stratified by Inflammation Status

Image Counts			
Split	**Active**	**Inactive/Minimal**	**Total**
Train	760	762	1522
Valid	217	218	435
Test	108	105	213
Total	1,085	1,085	2170
Patient Counts			
Split	**Active**	**Inactive/Minimal**	**Total**
Train	207	195	402
Valid	56	63	119
Test	22	34	56
Total	**285**	**292**	**577**
Lesion Counts			
Split	**Active**	**Inactive/Minimal**	**Total**
Train	297	305	602
Valid	83	93	176
Test	37	46	83
Total	**417**	**444**	**861**

Table 2. Training and Testing Performance Metrics for Deep Learning Models in Inflammation Classification*

Training Metrics				
Model	**Training accuracy**	**Validation accuracy**	**Training loss**	**Validation loss**
CNN	74.5%	79.8%	25.01	5.75
ResNet50 + 1 FC	82.5%	83.4%	6.09	1.79
MobileNetv2 100 + 2 FC	81.2%	81.8%	3.17	1.13
DenseNet121 + 3 FC	77.6%	80.9%	3.64	1.16
Testing Metrics				
Model	**Accuracy (95% CI)**	**Recall (95% CI)**	**Precision (95% CI)**	**F1 score (95% CI)**
CNN	80.2% (75.0-85.4)	70.85% (63.9-77.8)	87.55% (78.6-96.5)	77.85% (72.2-83.5)
ResNet50 + 1 FC	85.15% (81.7-88.6)	85.0% (79.7-90.3)	84.9% (76.7-93.1)	84.45% (79.7-89.2)
MobileNetv2 100 + 2 FC	85.9% (82.0-89.8)	90.5% (85.5-95.5)	82.05% (73.9-90.2)	85.7% (80.4-91.0)
DenseNet121 + 3 FC	83.4% (79.1-87.7)	76.15% (70.2-82.1)	88.85% (81.1-96.6)	81.65% (76.6-86.7)
ResNet50 + K-NN	89.7% (86.6-92.8)	87.35% (82.4-92.3)	90.95% (84.9-97.0)	88.85% (84.5-93.2)
MobileNetv2 100 + K-NN	86.65% (82.8-90.5)	84.95% (80.3-89.6)	87.7% (80.8-94.6)	86.0% (81.7-90.3)
DenseNet121 + K-NN	83.95% (80.6-87.3)	82.95% (78.2-87.7)	84.2% (76.5-91.9)	83.25% (78.6-87.9)

*The dataset includes both basal cell carcinomas (approximately 65%) and squamous cell carcinomas (approximately 35%). Images were captured at baseline, during active treatment (weeks 1-16), and post-treatment follow-up phases.
Abbreviations: CNN, convolutional neural network; FC, fully connected layer; K-NN, K-nearest neighbor; ResNet50, residual network with 50 layers; MobileNetV2, mobile network version 2; DenseNet121, dense network with 121 layers; F1, harmonic mean of precision and recall; CI, confidence interval.

RESULTS

•ResNet50 with K-NN achieved the highest test accuracy (89.7%, 95% CI 86.6-92.8), precision (90.95%, 95% CI 84.9-97.0), and F1 score (88.85%, 95% CI 84.5-93.2) (**Table 2**).
•K-NN approaches outperformed their linear probing counterparts by 0.55-4.55% in accuracy across all three feature extractors (**Table 2**).
•MobileNetV2 with linear probing achieved the highest recall (90.5%, 95% CI 85.5-95.5) (**Table 2**).
•The custom CNN showed the lowest accuracy (80.2%, 95% CI 75.0-85.4) but competitive precision (87.55%, 95% CI 78.6-96.5) (**Table 2**).

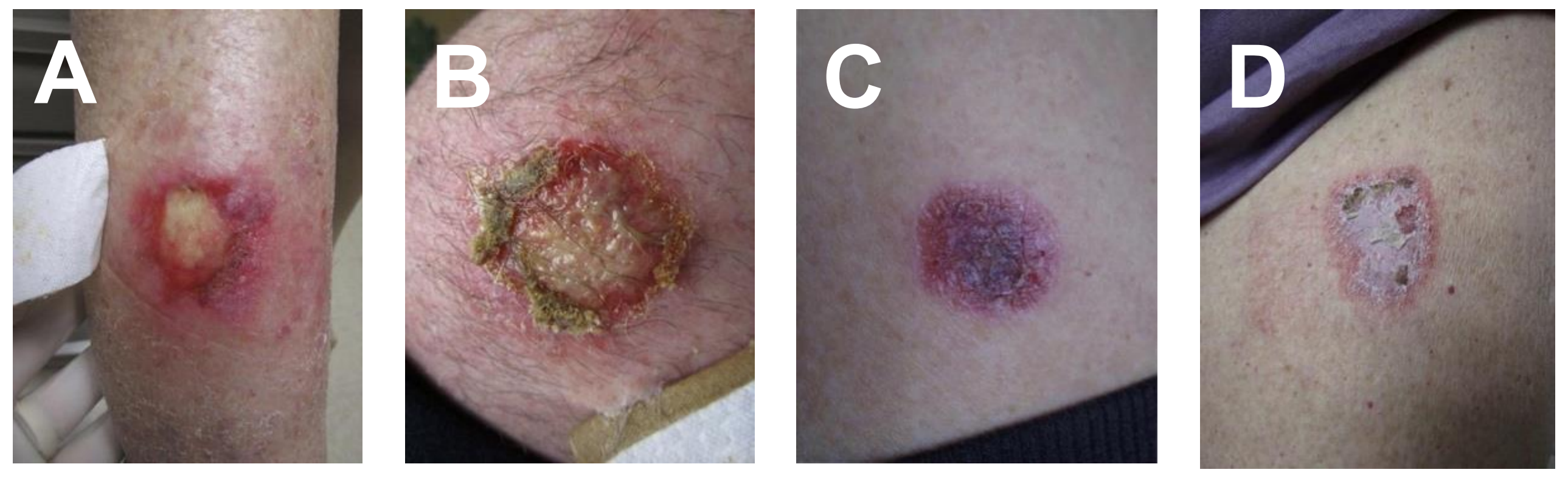

Figure 1. Representative Clinical Images of Active and Inactive/Minimal Inflammation Status in Keratinocyte Carcinomas Treated with Topical IMI/5-FU/RET. (A, B) Examples of "active" inflammation demonstrating intense erythema, scaling, and inflammatory response during treatment. (C, D) Examples of "inactive/minimal" inflammation showing mild residual erythema or post-inflammatory changes without active treatment reaction. ‘Active' images predominantly come from weeks 2-12 of treatment and 'inactive/minimal' images from baseline and post-treatment phases.

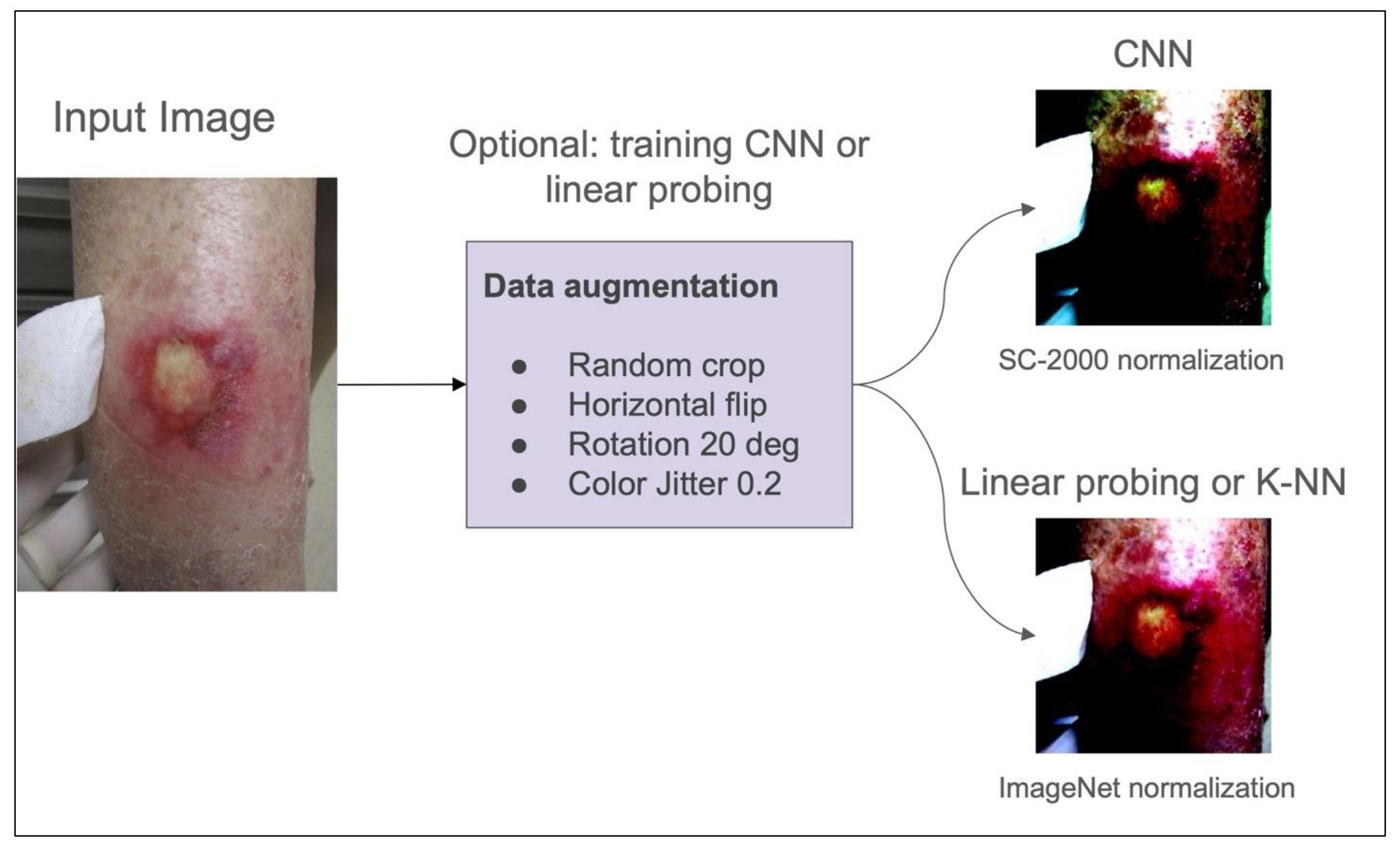

Figure 2. Data Preprocessing Pipeline for Deep Learning Classification of Keratinocyte Carcinomas. Schematic representation of the image preprocessing workflow used for DL classification of lesion inflammation status. All images were resized and cropped to 224 × 224 pixels. For CNN and feature extractor with linear probing, one of the following data augmentations was applied during training: resize and crop, horizontal flip, rotation by 20 degrees, or color jitter by 0.2. Images were normalized with mean and standard deviation from training data for CNN or from ImageNet for feature extractors.

DISCUSSION & CONCLUSION

•Pre-trained ResNet50 with K-NN effectively classifies KC inflammation status, offering a computationally efficient approach that requires no domain-specific training; K-NN's superiority over fine-tuning suggests ImageNet features optimally capture inflammation patterns, and additional training may degrade performance.
•This approach could standardize telemedicine monitoring and extend to field therapies for actinic keratoses and inflammatory dermatoses.
•Limitations: restriction to lighter skin (Fitzpatrick I-III), as IMI/5-FU/RET can cause prolonged hyperpigmentation in darker skin (Fitzpatrick IV-VI); the relatively small dataset also constrains CNN generalization.
•Clinical implementation requires external validation addressing skin tone diversity, image variability, and prospective assessment of clinical utility.