Interpreting Fine-grained Dermatological Classification with Deep Learning

S Mishra [1], H Imaizumi [2], T Yamasaki [1]

[1]The University of Tokyo

[2]ExMedio Inc

ISIC Skin Image Analysis Workshop

LONG BEACH
CALIFORNIA
June 16-20, 2019

exMedio

Scope

- Analyze model accuracy gap on benchmark datasets (CIFAR-10) vs. dermatological image corpus (DermAI*)
  - SOTA on CIFAR ~98%, whereas dermoscopic ~90%
- Investigate leading label pairs by case studies
  - 3 leading pairs investigated by GradCAM/GBP
- Suggestions on better datasets of user-submitted images by our experience
  - Data Augmentation, FoV, Gamma & Illumination correction

Dataset

User submitted Dermoscopic images across 10 most prevalent labels. 7264 images, split in 5:1 (train/test)

Acne

Alopecia

Blister

Crust

Erythema

Leukoderma

P. Macula

Tumor

Ulcer

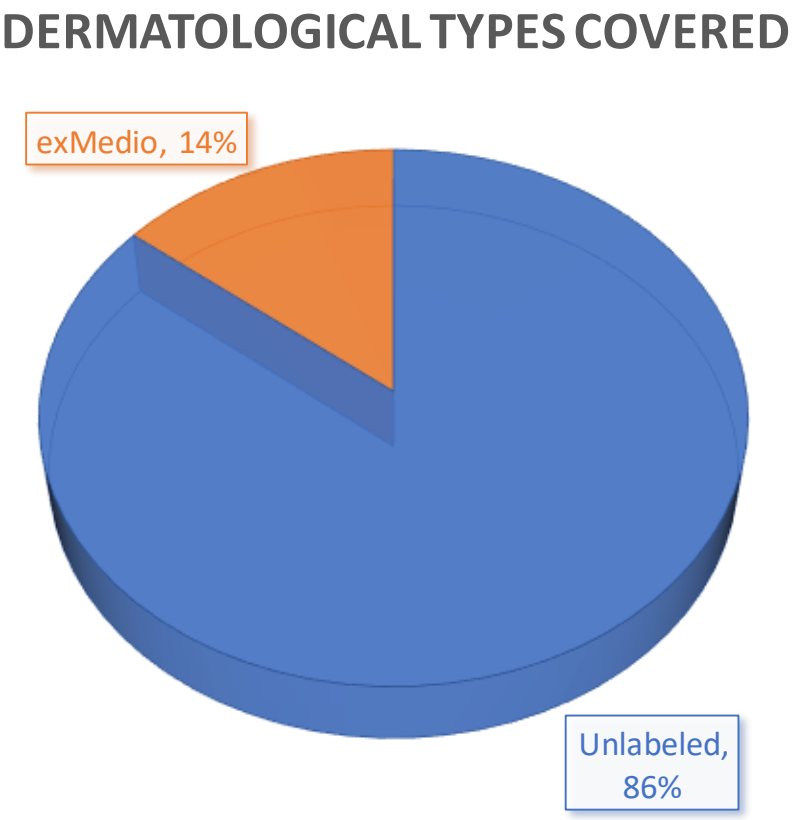
Wheal

Dataset

- Addressing the most common dermatological complaints.

- Ultimate goal:

To perform reliable rapid screening to reduce out-patient burden.



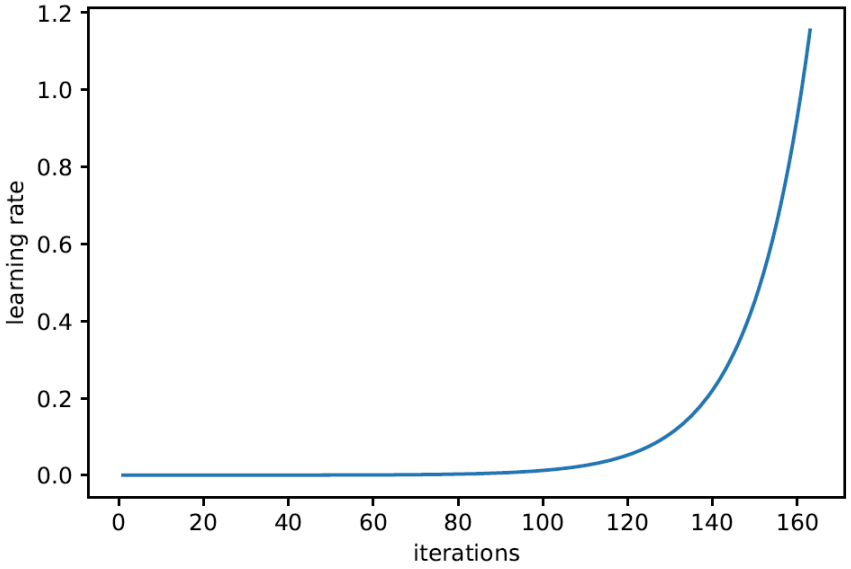
Model Learning

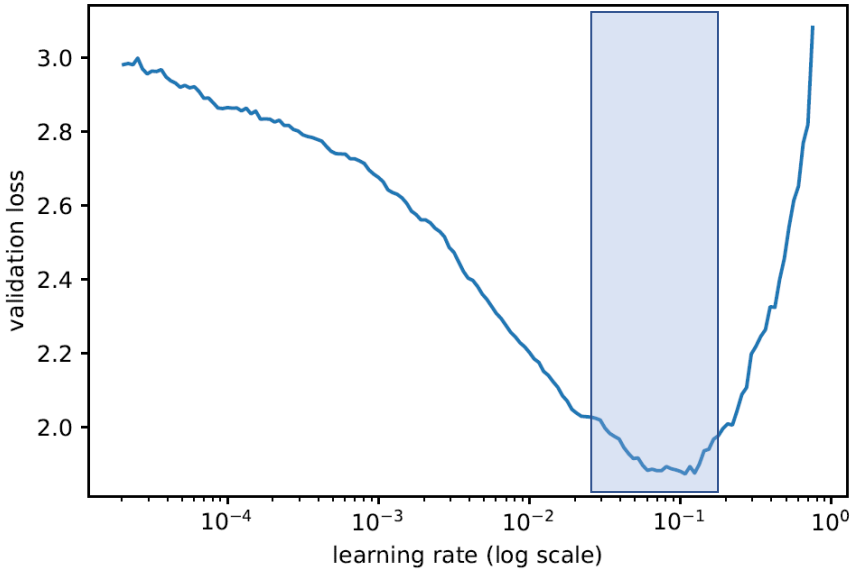
- Test several architectures of increasing size/complexity

  Resnet-34, ResNet-50, ResNet-101, ResNet-152

- 5:1 split, Early stopping, BCE with logits loss
  - Learning rate range test
  - SGD + Restarts (SGD-R)
  - SGD-R + Length Multiplication+ Differential Learning
- Modus operandi tested on CIFAR-10 prior*

Learning Rate range-test





Steadily increase the LR and observe the Cross entropy loss

Test several mini-batches to see a point of inflexion

Reference:

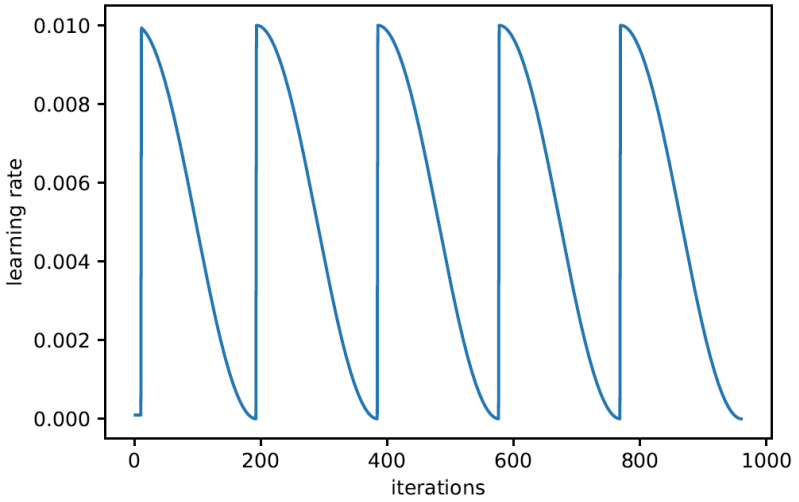
Cyclical Learning rates for training NN, L. Smith [2017]

Deep Learning, S. Verma et al. 2018

SGD-R

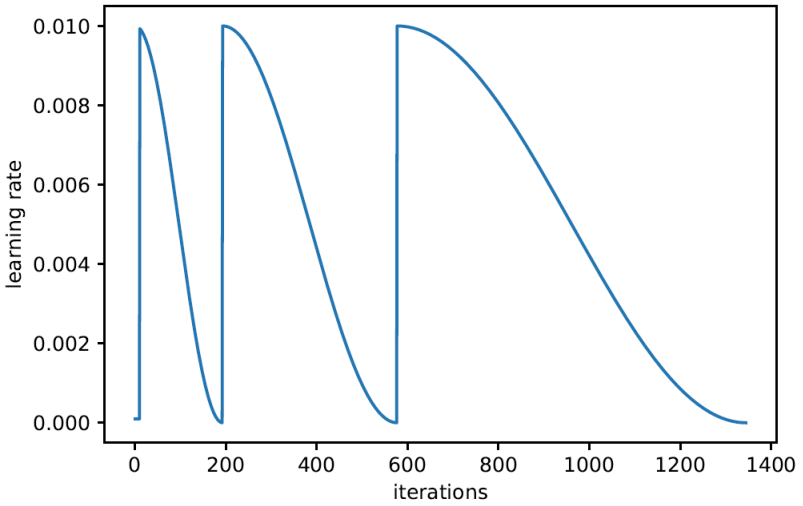
1. Avoid monotonicity by Cosine scheduling function

$$v(t) = \frac{1}{2}\left(1 + v\, cos\left(\frac{t\pi}{T}\right)\right) + \varepsilon$$



Initial coarse fit by tuning the last (or last few) FC layer

2. Cycle Length Multiply by integral powers of 2 over whole architecture

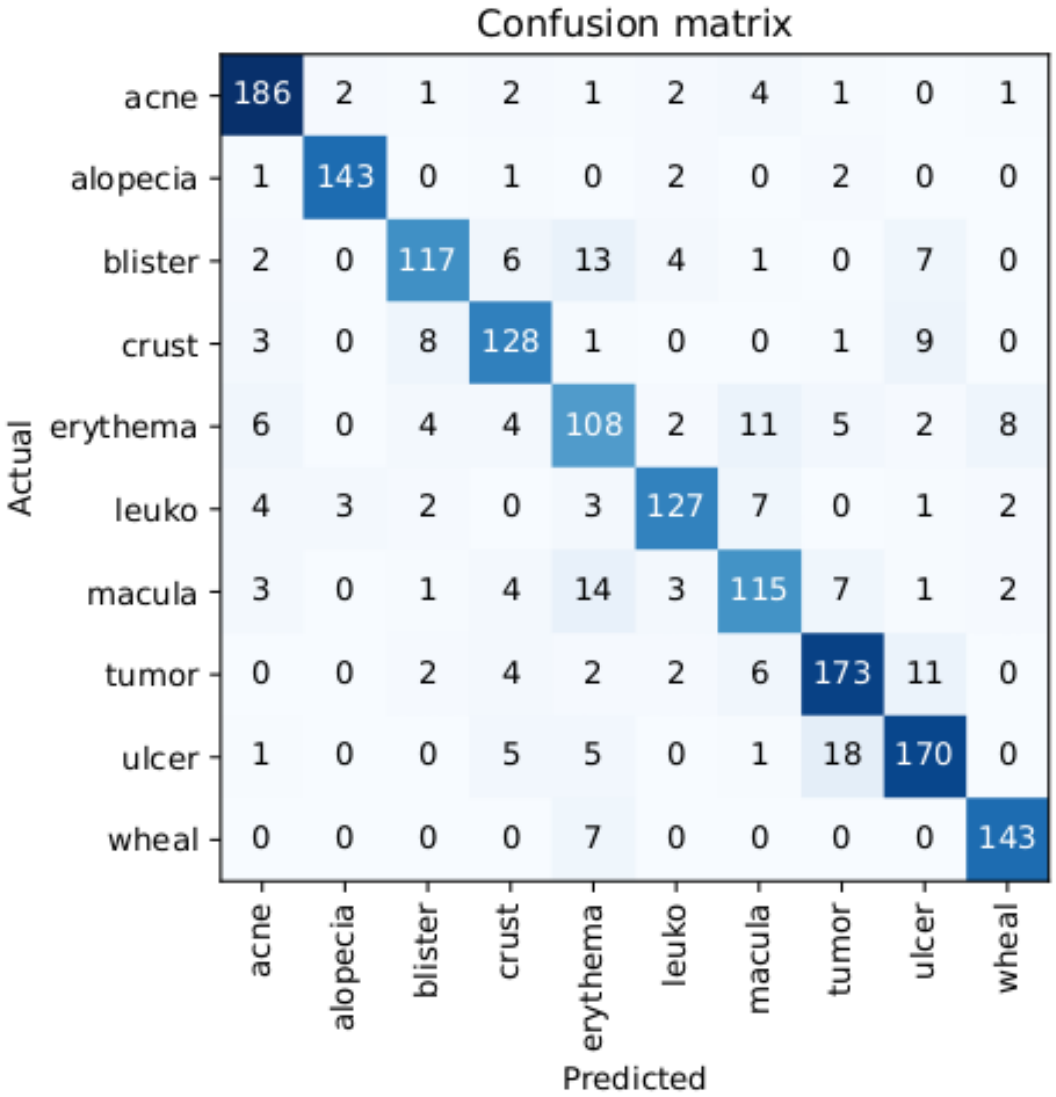


Tighter fit over all layers

Reference:
SGD with Warm restarts, Loschilov [2017]

Application

Architecture	Acc. (Top-1)
ResNet-34	88.9%
ResNet-50	89.7%
ResNet-101	88.2%
ResNet-152	89.8%



ResNet 152 Confusion Matrix

Analysis

- Following best practices still leaves gap.
- Focus on the label pairs which account for most errors.
- Use GradCAM and Gradient Backprop to analyze what CNNs capture in learning process.

Label 1	Label 2	Counts
Ulcer	Tumor	29
Macula	Erythema	25
Blister	Erythema	17
Erythema	Wheal	15
Crust	Ulcer	14
Blister	Crust	14
Macula	Tumor	13
Macula	Leukoderma	10
Blister	Ulcer	7
Tumor	Erythema	7
Crust	Tumor	5

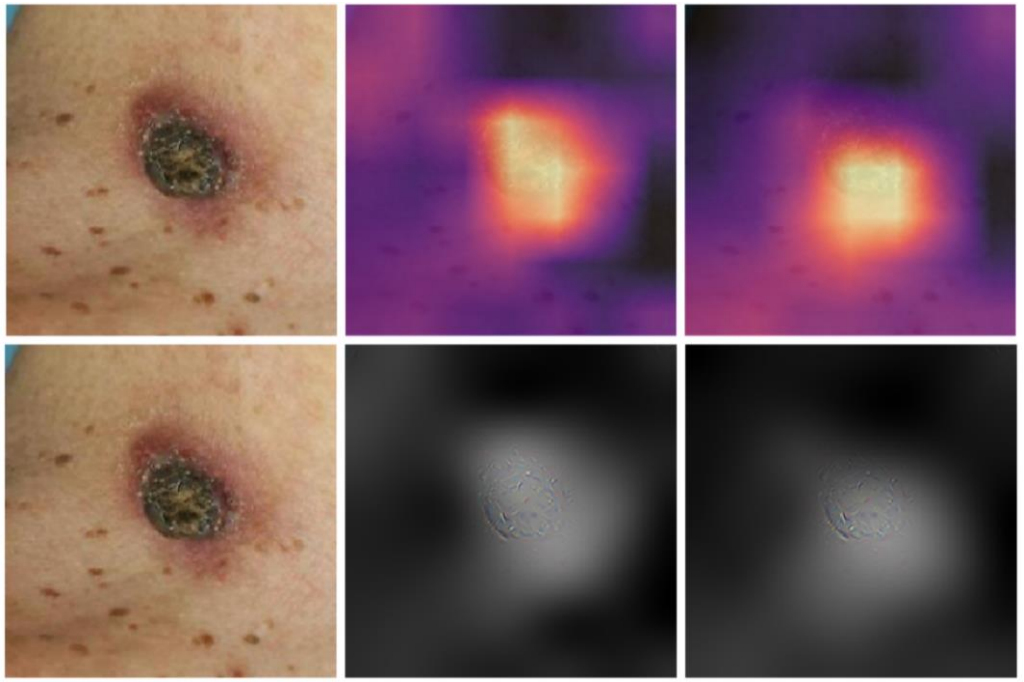
Label pairs with at least 5 errors

Reference:
GradCAM: Visual explanation from DNN, Selvaraju [2016]
Guided BP, Springenberg [2014]

Ulcers & Tumors

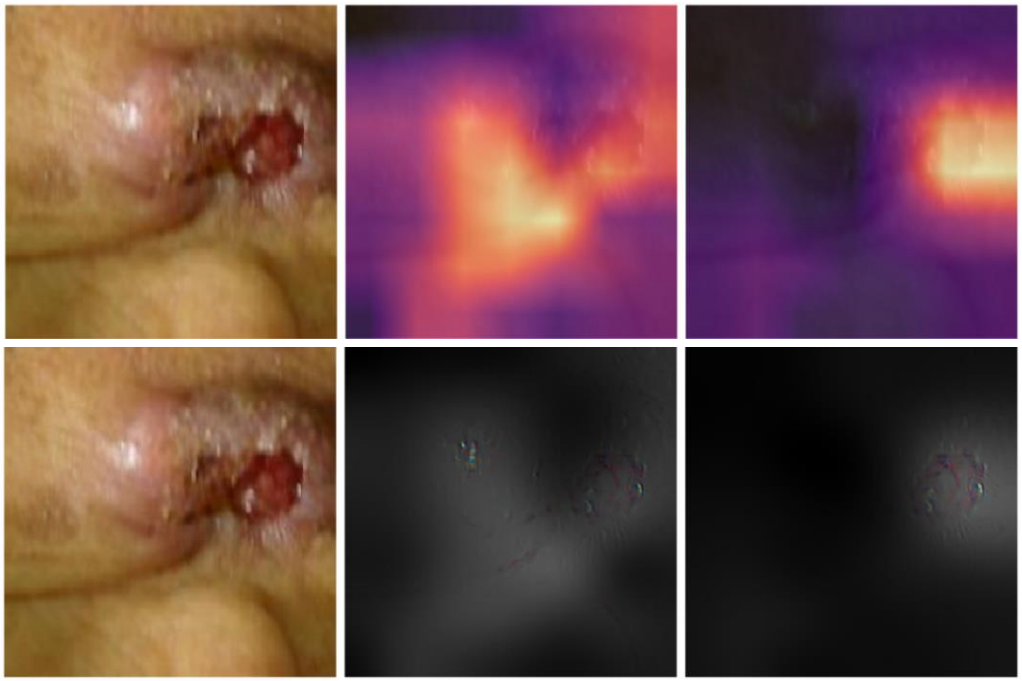
Ulcer 0.391	Tumor 0.152

High degree of geometrical (spherical) similarity is the common factor in many samples



Tumor 0.78	Ulcer 0.212

Elevations and inflammations seen in Tumors, misclassifies many ulcer samples.



Macula & Erythema

Erythema 0.53	Macula 0.41

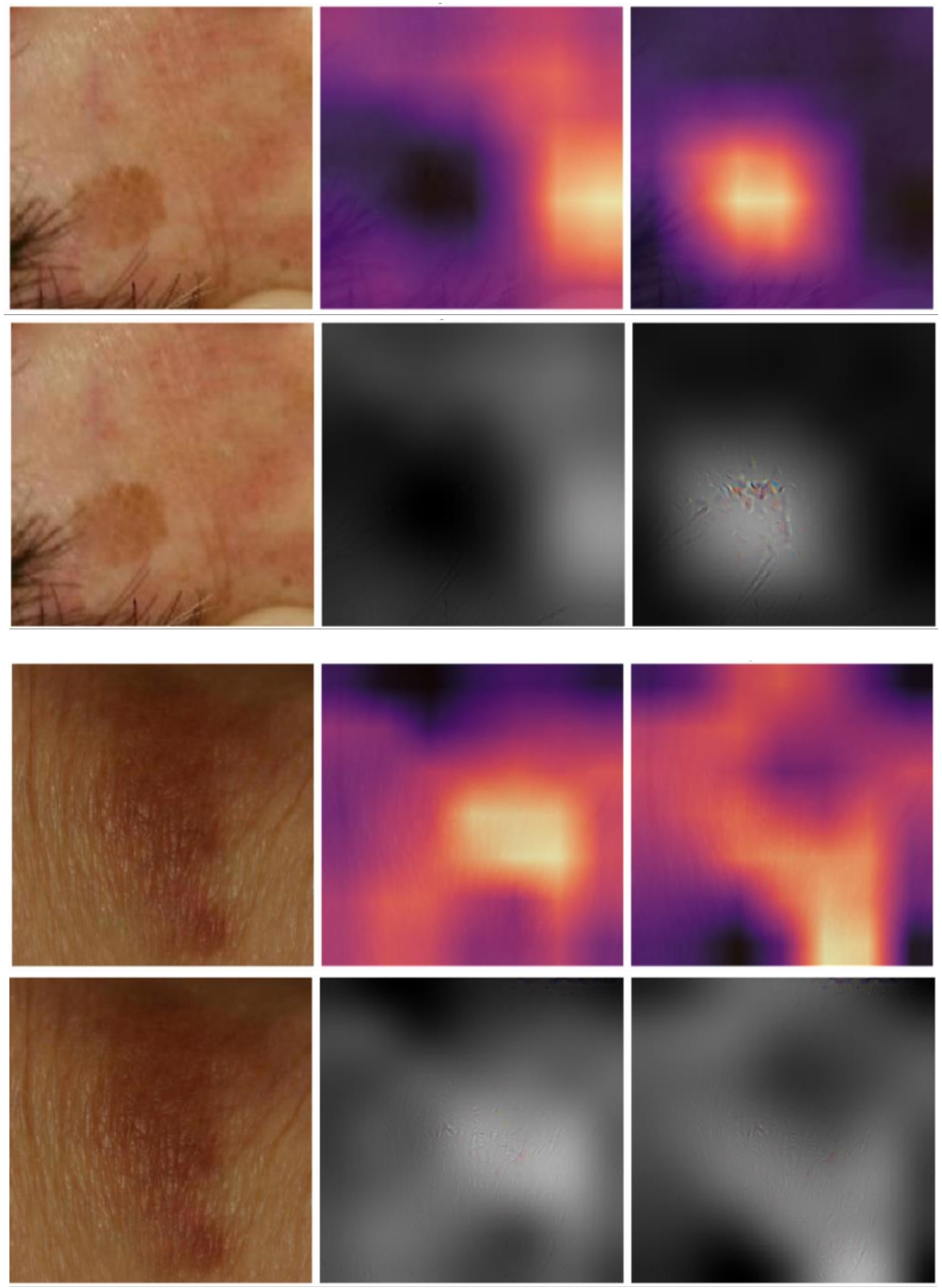
Presence of pigmentation patches around the lesion can mispredict.

FoV and ROI selection could lead to better results.

Macula 0.69	Erythema 0.28

Oval/cycloidal patches makes GBP confused with the overall shape of Macula.

FOV & Depth important factors to consider



Ulcer & Crust

Crust 0.86	Ulcer 0.124

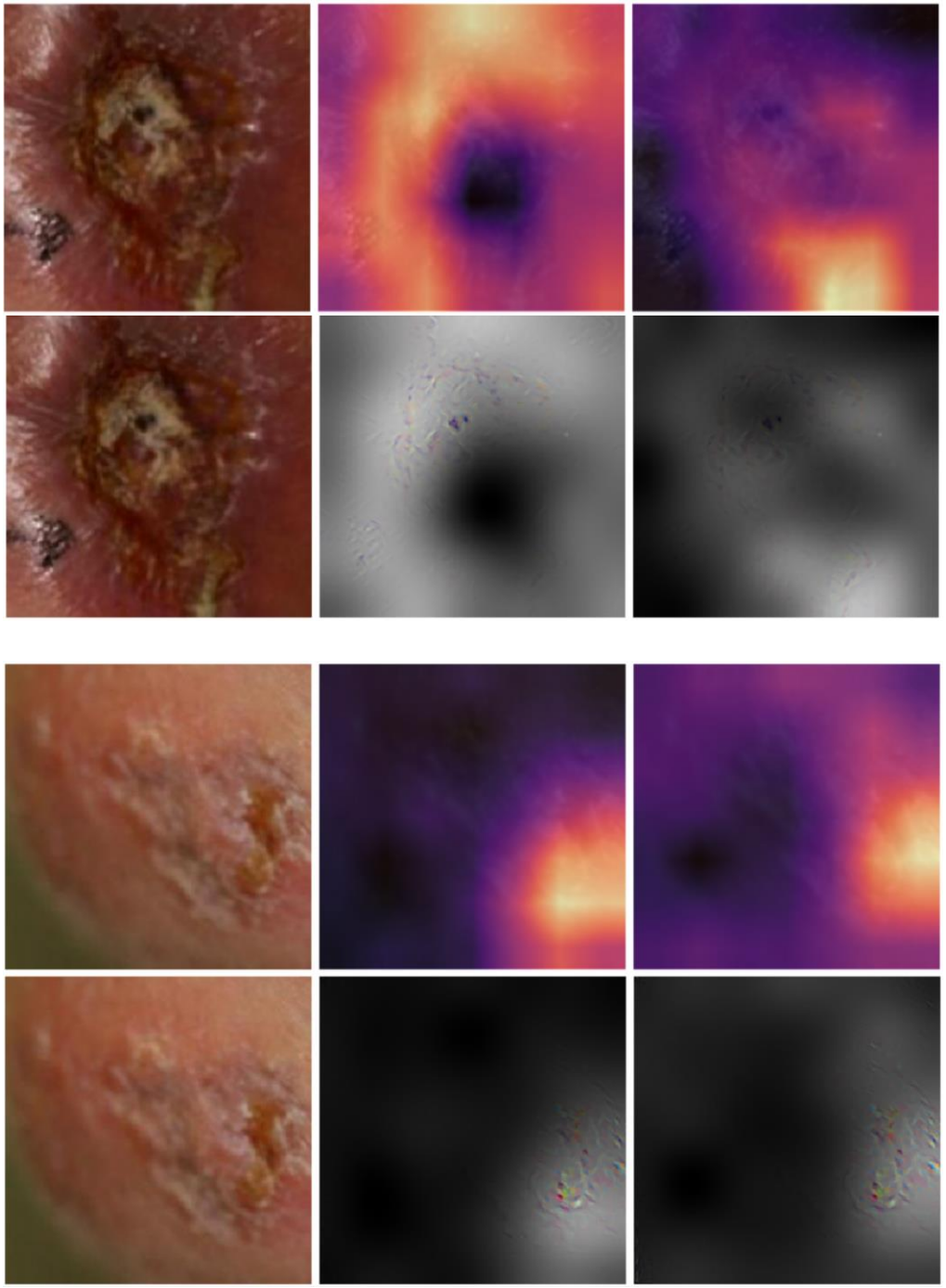
Presence of large centroid is possible source.

Difficult to predict as both related chronologically

Ulcer 0.91	Crust 0.06

Oval/cycloidal patches on GBP

Selection of right RoI, illumination could improve many cases.



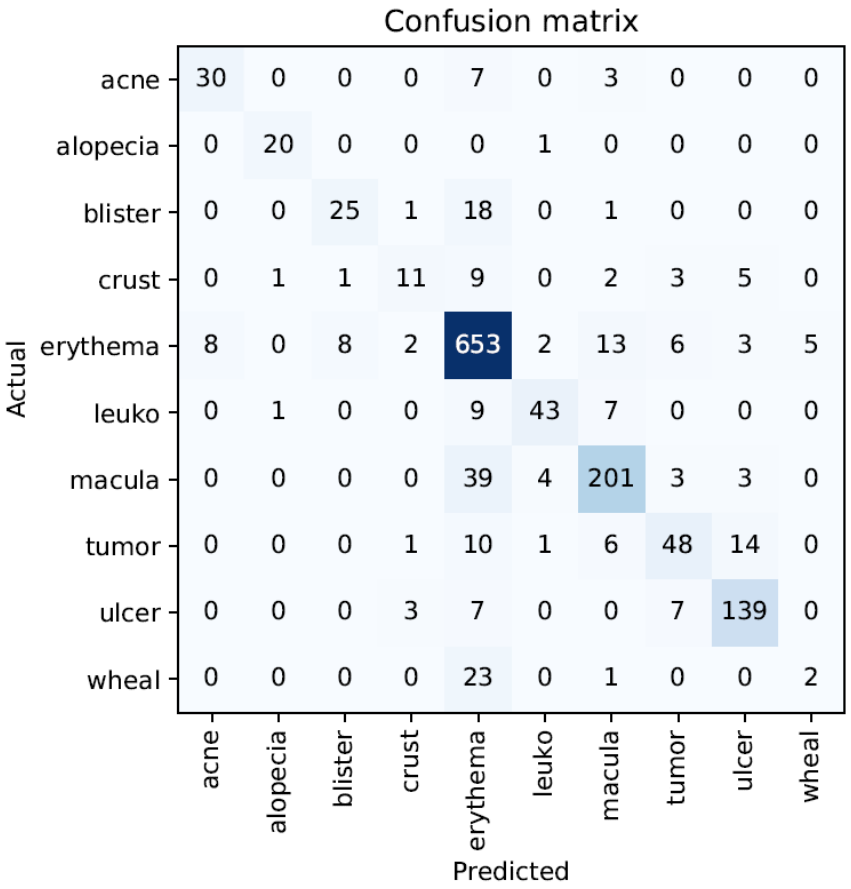
Mitigation

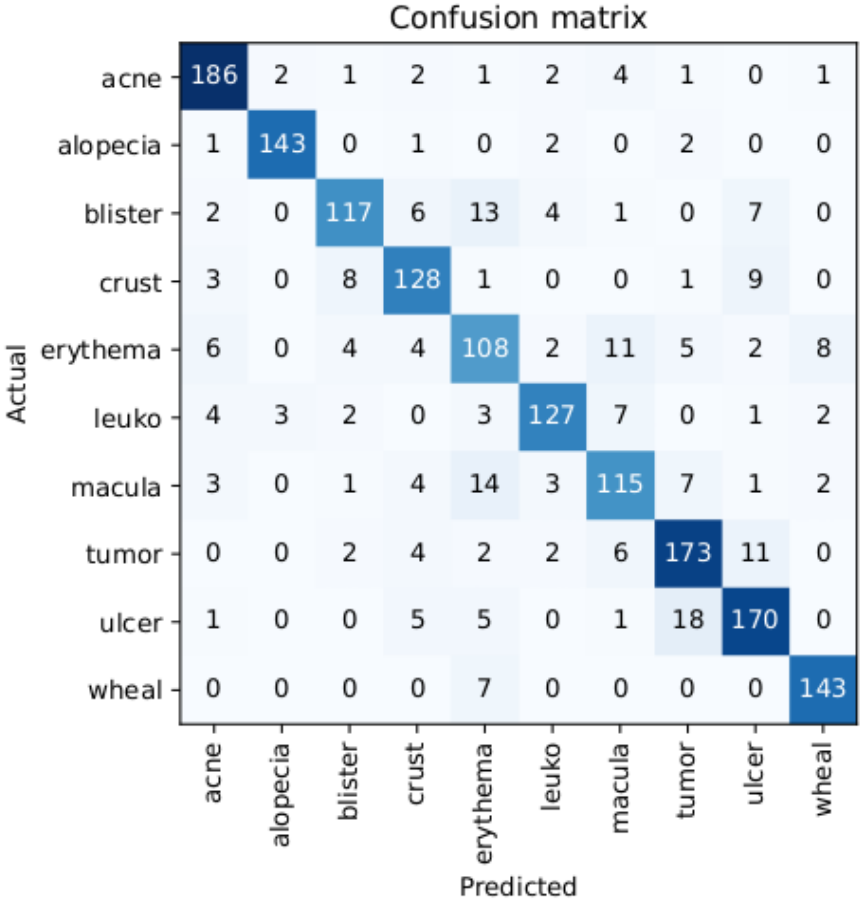
Highlight some of the “hard-learned lessons” building this project from scratch.

Mitigation factors to look out:

- Balancing training sets (dynamic vs static)
- Field of View / ROI selection
- Illumination and Gamma correction

Balancing for model learning



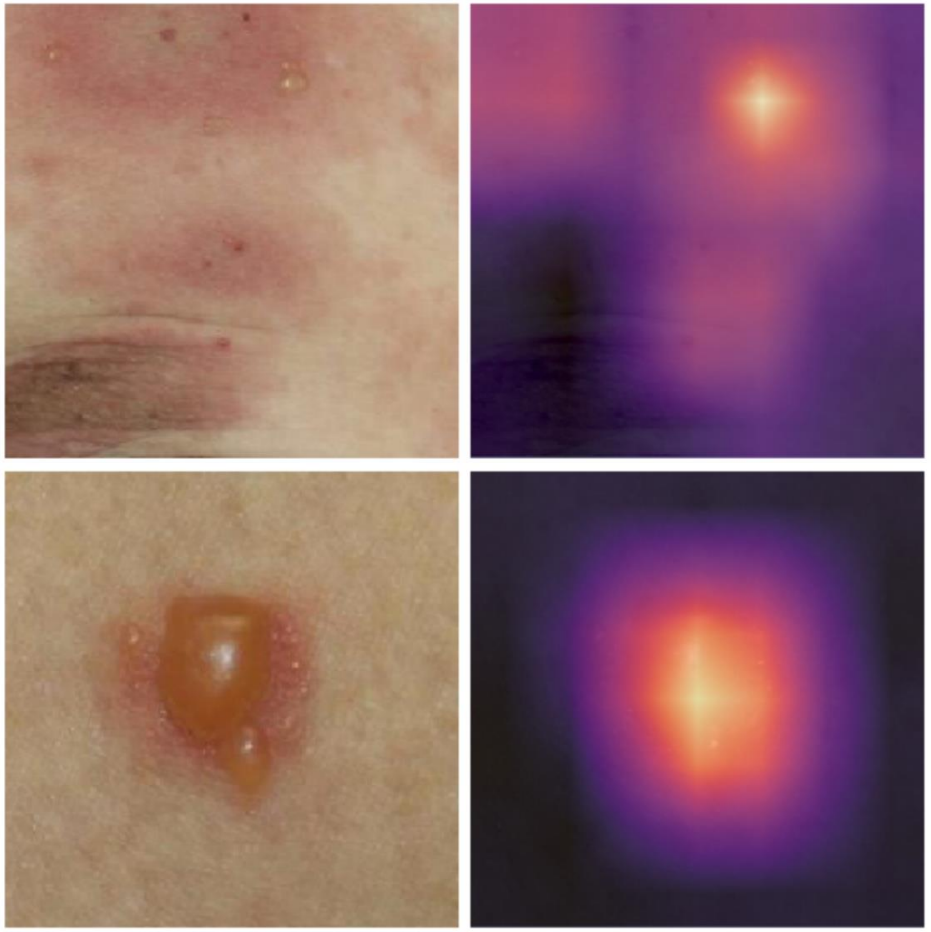


Custom datasets can be small, unevenly divided. Best to use dynamic in-memory augmentation during batch selection. Larger batches preferably.

Field of View/Object Depth

P [Blister]	0.547

P [Blister]	1.000

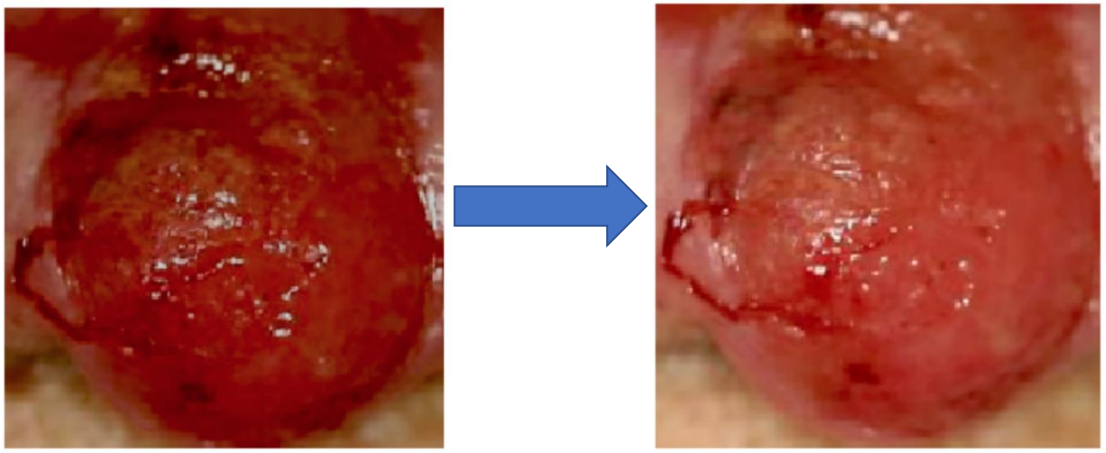


FOV selection dramatically improves performance. In user-submitted images, pre-processing needed. Bonus: if illumination stable

Gamma & Illumination

Often illumination & shadow effects

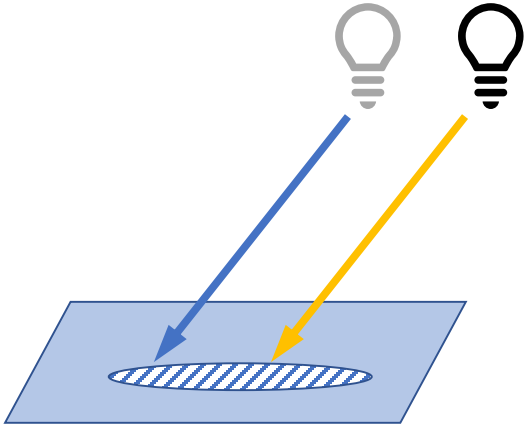
Gamma adjustment ≈ 1.2 – 1.5

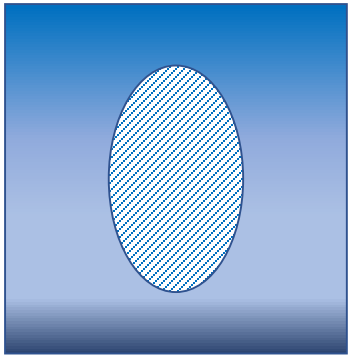


Prediction : Ulcer (98%)
Actual : Tumor (1%)

Prediction : Tumor 78%

Creating illumination map & reversing imbalanced lighting by normalizing.





Conclusion

- Gap may never be entirely removed,
- [Status Quo] Racial diversity one of the hardest problems to crack. Better to focus on single one for better performance. (But harder in developed countries).
- Not all artifacts can be fixed in user-submitted images.
- Augmentation & Photo-grammatic corrections can improve the quality of model learning/inference dramatically.
    - Balancing training data, FOV reduction, Gamma & illumination correction

https://github.com/souravmishra/ISIC-CVPRW19

Thank you!